The Dissertation Committee for Kishore John Doshi Certifies that this is the approved version of the following dissertation:

RNA Secondary Structure Prediction and an Expert Systems Methodology for RNA Comparative Analysis in the Genomic Era

Committee:

Robin Gutell, Supervisor

Rick Russell

Edward M. Marcotte

David E. Graham

Dimitrii E. Makarov

RNA Secondary Structure Prediction and an Expert Systems Methodology for RNA Comparative Analysis in the Genomic Era

by

Kishore John Doshi B.S.; M.S.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

May 2007

Dedication

To all my friends and family who have provided much inspiration and encouragement during what has undoubtedly been the most difficult time period in my life. On numerous occasions, they have exhibited more confidence in my abilities than I have.

Acknowledgements

In September 2001, I began a long journey to transform myself from a software consultant at Cap Gemini Ernst and Young into a hybrid of a biologist, computer scientist and an engineer. This dissertation marks the end of the first phase in that journey. Over the last six years, I have been fortunate to work with many different talented, dedicated and conscientious individuals and I would like to take a moment to acknowledge them here.

I must begin my thanking Dr. Robin Gutell for taking a significant risk and providing me with an opportunity to work in his research lab, engaging in many long, thought provoking conversations and acting as my biggest advocate. I wish to thank Jamie Cannone for his guidance and support, and Dr. Jung Lee, for many through provoking conversations on many diverse topics including RNA structure and backpacking through Germany. Dr. Lee, Jamie and Henriette Ries have provided invaluable feedback on the software toolkit I developed for RNA comparative analysis, CAT. I wish to thank Dr. Gutell's systems administrator, Craig Dupree for keeping the computer systems in the lab in top condition and going above and beyond to assist me with any issues, including working with Windows Server 2003 which is a constant source of frustration for Craig being a UNIX god. I must also thank Tim Guinn, who was the systems administrator in the Gutell Lab before Craig. I would like to thank David

Gardner, Stuart Ozer of Microsoft Research, Dr. Phil Cannata of Sun Microsystems, and Dr. Mike Keys of Sun Microsystems who have supported my research efforts and collaborated with me for separate periods of time over the last three years on the development of novel database solutions for RNA comparative analysis. I must also thank Dr. John Boisseau and his staff members from the Texas Advanced Computer Center (TACC), specifically Tomislav Urban and Shirley Cohen who have been advocates of my work and have supported my specific research efforts. I would like to thank Dr. David Hillis for supporting my research efforts through the IGERT program. Finally, I would like to thank the members of my committee, Dr. Rick Russell, Dr. David Graham, Dr. Edward Marcotte and Dr. Dimitrii Makarov.

On the personal side I must thank my parents, Jeanne and Kishore Doshi who have always believed in me and supported my efforts. Furthermore, I must thank my closest friends who have been a significant source of encouragement over the last six years especially on the numerous occasions on which I have considered quitting, Edison Morales, Christy Howard, Dean Walser, Joe Sayakumane, and Shaji Georgekutty.

RNA Secondary Structure Prediction and an Expert Systems Methodology for RNA Comparative Analysis in the Genomic Era

Publication No._____

Kishore John Doshi, Ph.D.

The University of Texas at Austin, 2007

Supervisor: Robin Gutell

The ability of certain RNAs to fold into complicated secondary and tertiary structures provides them with the ability to perform a variety of functions in the cell. Since the secondary and tertiary structures formed by certain RNAs in the cell are central to understanding how they function, one of the most active areas of research has been how to accurately and reliably predict RNA secondary structure from sequence; better known as the RNA Folding Problem. This dissertation examines two fundamental areas of research in RNA structure prediction, free energy minimization and comparative analysis. The most popular RNA secondary structure prediction program, Mfold 3.1 predicts RNA secondary structure via free energy minimization using experimentally determined energy parameters. I present an evaluation of the accuracy of Mfold 3.1 using the largest set of phylogenetically diverse, comparatively predicted RNA secondary structures available. This evaluation will show that despite significant revisions to the energy parameters, the prediction accuracy of Mfold 3.1 is not significantly improved when compared to previous versions. In contrast, RNA comparative analysis has

repeatedly demonstrated the ability to accurately and reliably predict RNA secondary structure. The downside is that RNA comparative analysis frequently requires an expert systems methodology which is predominately manual in nature. As a result, RNA comparative analysis is not capable of scaling adequately to be useful in the genomic era. Therefore, I developed the Comparative Analysis Toolkit (CAT) which is intended to be the fundamental component of a vertically integrated software infrastructure to facilitate high-throughput RNA comparative analysis using an expert systems methodology.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.A Background Summary

It was first established in 1959 using physical chemical experiments that Tobacco Mosaic Virus RNA could form a partial helical structure[1]. Subsequently it was shown that other RNAs such as Ribosomal RNA (rRNA) and Transfer RNA (tRNA) were also capable of forming partial helical structure[2]. As a result of these experiments, a model was proposed for RNA secondary structure which consisted of DNA style helices of a minimum length of four base pairs formed from intramolecular interactions between nucleotides of the RNA chain[3]. Furthermore, the model speculated that the partial helical structure observed in the physical chemical experiments was achieved through "looping out" nucleotides on both strands of the helix[3]. Since this initial secondary structure model was postulated, it is now well-understood that certain RNAs are capable of forming extremely complicated tertiary structures which includes a secondary structure that consists of DNA-style helices with canonical base pairs (G:C, A:U and G:U) as well as non-canonical base pairs and conformations.

The ability of certain RNAs to fold into complicated secondary and tertiary structures provides them with the ability to perform a variety of functions in the cell. It is now known that RNA plays a role fundamental role in protein synthesis[4-10] and the self splicing of Group I introns[11-13]. RNAs have been implicated in the regulation of gene expression through riboswitches[14, 15] and micro-RNAs[16]. Riboswitches are highly structured, ligand binding domains found in the non-coding regions of mRNA transcripts[14]. The expression of an mRNA which contains a riboswitch is regulated in response to the presence or absence of a small molecule which binds in the domain[14]. Micro-RNAs are encoded in the genome and start out as pre-miRNAs which form a

characteristic secondary structure[16]. Pre-miRNAs are subsequently processed into single stranded RNAs of *c.a.* 22 nucleotides which can post-transcriptionally silence the expression of a gene through the RNA interference pathway[17-19]. It has recently been shown that altered expression of specific miRNA genes contributes to the initiation and progression of cancer[20-23].

Since the secondary and tertiary structures formed by certain RNAs in the cell are central to understanding many aspects of Biology, one of the most active areas of research has been how to accurately and reliably predict an RNA secondary structure from its sequence; better known as the RNA Folding Problem. The problem is enormously complex. Consider the 16S Ribosomal RNA (rRNA) from Escherichia coli which has 1542 nucleotides. The number of possible secondary structure helices of length four or more is 14,684 and number of possible secondary structure models that can be assembled from those helices is *c.a.* 4.3×10^{393}[24]. In contrast, the comparatively predicted secondary structure model for Escherichia coli 16S rRNA has only 58 helices[25]. In order to reduce the complexity of the problem, constraints or knowledge must be introduced. Two different methodologies have been predominantly used to predict RNA secondary structure from its sequence. One methodology, free energy minimization, involves the identification of the minimum free energy secondary structure for an RNA sequence based on experimentally determined sequence-dependent energetic parameters. The second, RNA comparative analysis, relies on the identification of common patterns of conservation and variation from a large set of homologous RNA sequences which are assumed to have a common structure.

These two RNA secondary structure prediction methods have important similarities and differences. Free energy minimization is based on experiments on small RNA oligoribonucleotides which are then applied to any RNA regardless of size.

Because the number of possible oligoribonucleotides is so large, it is not possible to conduct experiments for all of them. As a result free energy minimization is only capable of predicting secondary structure base pairs with Watson-crick exchanges (G:C, A:U, G:U). Furthermore, the energetics of many complicated loop structures observed in different RNA secondary structures are only be estimated. In contrast, comparative analysis is a knowledge-based approach. Comparative analysis is based on the assumption that homologous RNAs from different organisms share a common structure. By collecting homologous RNA sequences from many different organisms, the common structure can be deduced. Methods for deducing the common structure of a set of homologous RNAs make no assumptions about the type of RNA structure formed. Comparative analysis is capable of deducing secondary structure base pairs with Watson-crick exchanges as well as non-canonical base pairs (*e.g.,* U:U, C:C). Comparative analysis can accurately predicted complicated loop structures and even some tertiary interactions. Finally, comparative analysis can be applied to sets of homologous RNAs regardless of size without any concern about extrapolation.

The free energy minimization approach to RNA secondary structure prediction is grounded in the basic principles of physical chemistry. These principles dictate that an RNA secondary structure which forms spontaneously must represent a free energy minimum. Therefore, if one could develop a thermodynamic model which adequately reflects the complexities associated with RNA secondary structure, then that model could be used to predict the biologically relevant RNA secondary structure for a given sequence by simply identifying the minimum free energy structure. The nearest-neighbor thermodynamic model for RNA secondary structure[26, 27] was initially based on the concept that the single largest factor in the stability of a secondary structure helix relative to the single strand was the energy stabilization due to stacking interactions between the

individual base pairs[28]. Melting experiments were conducted using short oligoribonucleotides to measure the effect of sequence dependence on the stability of different RNA duplexes[29-31]. Using these experimentally determined energy parameters and the nearest-neighbor model, the free energy of a given RNA structure could be calculated. Subsequently, the first programs were developed to identify the minimum free energy structure for a given RNA sequence using dynamic programming techniques[32-34].

Since this initial foundation was developed, a significant amount of research has focused on: 1) improving the sequence dependence of the energy parameters and 2) measuring parameters for different loop structures observed in RNA secondary structure such as internal loops or bulge loops[35-37]. The updated energy parameters were included in different RNA folding programs such as Mfold 2.3[37, 38] or RNAfold[39]. The accuracy Mfold 2.3 was rigorously evaluated using a diverse set of 16S and 23S rRNA secondary structures predicted with comparative analysis and reported to have an average prediction accuracy of 46% for 16S rRNA and 44% for 23S rRNA[40, 41]. The prediction accuracy for an individual 16S or 23S rRNA sequence could be as high as 80% or as low as 20%[40, 41]. After this evaluation was published, the sequence dependence of the energy parameters was expanded and Mfold 3.1 was released[42-44]. An evaluation of Mfold 3.1 by Mathews et al indicated that an average about 73% of the known secondary structure base pairs for a given RNA were predicted correctly for sequences up to 700 nt, and an average of 97.1% of the known secondary structure base pairs were observed if one considered the population of the first 749 suboptimal structure predictions with up to 20% higher free energy than the minimum energy prediction[44].

In contrast to free energy minimization, which is a physical chemical approach to predict RNA structure based on experimentally determined energetic parameters, RNA

comparative analysis is a knowledge-based approach for predicting RNA structure through the analysis of a diverse set of homologous RNA sequences under the assumption that they form a common structure. The original tRNA secondary structure was determined manually through comparative analysis[45, 46] and later verified by high resolution X-ray crystallography[47, 48]. In 1975, Fox and Woese manually determined the secondary structure of the 5S rRNA through the identification of phylogenetically conserved helices[49]. In 1980 the first 16S and 23S rRNA secondary structures were postulated using alignments of two sequences respectively and the manual identification of *positional covariations*[50, 51]. *Positional covariations* were defined as columns in the alignment that exhibited coordinated, compensating changes between canonical base pairs (i.e., G:C <=> A:U). As the number of RNA sequences available increased, more systematic covariation analysis techniques were developed to systematically identify *positional covariations* in alignments of homologous RNA sequences without any built-in biases about expected base pair types[52, 53]. Covariation analysis techniques were used to enhance the 16S and 23S rRNA secondary structures models and contributed to the identification of many novel RNA structural characteristics and motifs such as: non-canonical base pairs, pseudoknots, parallel helices and tetraloops[52, 54, 55]. Throughout this period, databases of comparatively predicted 16S and 23S rRNA comparative structure models were made available via the internet[56-59]. In 2000, the first high resolution X-ray crystal structures became available for 16S and 23S rRNA, validating the comparatively predicted structure models and the comparative analysis methodology[25, 60, 61]. Several additional RNA molecules have been studied from a comparative perspective including: group I Introns[11, 62-64], group II Introns[65, 66], RNaseP[67, 68], telomerase RNA[69, 70], tmRNA[71], U RNA's[72], SRP RNA [73],

various untranslated regions (UTR) of mRNAs such as RNaseE[74] and the T Box transcription antitermination system[75]

1.B ORGANIZATION OF THIS DISSERTATION

This dissertation is organized around two fundamental areas of research in RNA structure prediction, free energy minimization and comparative analysis. In Chapter 2, I present an evaluation of Mfold 3.1[44] using the largest set of phylogenetically diverse, comparatively predicted RNA secondary structures available. The comparative structure database used for the evaluation contains Ribosomal RNA (5S, 16S and 23S) as well as Transfer RNA and is well-balanced between longer and short RNAs. The goal of this evaluation is not to simply determine the accuracy of Mfold 3.1, but to fundamentally question if the nearest-neighbor model can adequately represent the complicated arrangement of secondary structure helices observed in known RNA structures. Is solving the RNA Folding problem just a matter of determining more sequence dependent energy parameters? Is the most important factor in the energetic stability of an RNA structure the formation of consecutive base pairs in a helix? Should an RNA secondary structure be predicted using a dynamic programming algorithm which initially considers the formation of base pairs and helices independent of their arrangement? While there is no debate that the nearest neighbor thermodynamic model is fundamentally relevant for describing RNA secondary structure, does it need to be expanded?

In this evaluation, the top 749 suboptimal structure predictions as well as the minimum free energy secondary structure prediction are investigated for each RNA sequence in the comparative structure database. RNA sequences are folded with Mfold 3.1 as complete sequences, and the accuracy of a minimum free energy secondary structure predicted by Mfold 3.1 is tested against the comparatively predicted secondary structure. The secondary structure predictions and accuracies of Mfold 3.1 are compared

with its predecessor, Mfold 2.3[37, 38], and all metrics considered in the evaluation of Mfold 2.3[40, 41] are revisited in this evaluation. A new evaluation metric which considers the prediction accuracy of individual base pairs as a function of the number of intervening nucleotides between the 5' and 3' halves of the base pair is introduced, and the complete set of suboptimal secondary structure predictions for all 16S rRNAs in the comparative structure database are characterized for prediction accuracy as well as the total number of comparative base pairs observed throughout the population.

In Chapter 3, I focus on RNA comparative analysis. While RNA comparative analysis has demonstrated the ability to accurately and reliably predict complicated RNA secondary structures, the analysis has traditionally been conducted manually. The CRW Project[62] was established to identify RNA sequences of interest in public sequence repositories such as GenBank[76] including 5S, 16S, and 23S rRNAs and tRNA; analyze them by comparative analysis; and disseminate the results to the scientific community. The CRW Project has developed its own software infrastructure to facilitate manual RNA comparative analysis using an expert systems methodology. In light of the significant advances in sequencing technology as a result of the human genome project, the number of RNA sequences of interest to the CRW Project has grown exponentially since 2002, and the CRW Project has been unable to analyze all these sequences using their expert systems methodology. Therefore, it was necessary to scale up the expert systems methodology originally developed by the CRW Project and in the progress create a vertically integrated software infrastructure for RNA comparative analysis that: 1) provides for significant quality control and 2) is capable of leveraging disparate, non-connected data sources. The first step in the development of this infrastructure was the development of the Comparative Analysis Toolkit (CAT), a new suite of bioinformatics tools specifically designed to complement the expert systems methodology for RNA

comparative analysis used by the CRW Project. The development of CAT began in the fall of 2003, and was first released for analysis work within the CRW Project in early 2004. Since its initial release, I have been continually enhancing CAT by developing tools for semi-automatic RNA sequence alignment and automated RNA sequence alignment evaluation. CAT is directly integrated with other resources of the CRW Project such as the RNA metadata database and has significantly increased the ability of the CRW Project to analyze RNA sequences of interest via comparative analysis techniques.

Chapter 2: RNA Secondary Structure Prediction via Free Energy Minimization

2.A INTRODUCTION

In Chapter 1, I discussed the basic concept that certain RNAs are capable of spontaneously folding into a secondary structure in-vivo that consists of DNA style helices and single stranded regions[1-3]. Base pairs in these helices are the result of intramolecular interactions between different nucleotides on the given RNA chain. Shortly after the characteristics of RNA secondary structure were determined, a thermodynamic model for RNA structure was postulated, the nearest-neighbor mode[26, 27]. This model was based on the concept that the coaxial stacking of base pairs was the most important factor in the stability of an RNA secondary structure helix[28]. By 1974, melting experiments were conducted with short oligoribonucleotides to determine the sequence dependence on the stabilities of different RNA duplexes[29-31].

By 1987, the computer program Mfold was developed to predict the secondary structure for any RNA sequence. Mfold utilized an efficient dynamic programming algorithm for identifying the minimum free energy secondary structure based on sequence thermodynamic parameters experimentally determined for RNA duplexes [35]. In the intervening time period between 1987 and 1999, the thermodynamic energy parameters and the dynamic programming algorithm were both areas of active research. Research into the thermodynamic parameters focused on: 1) expanding the sequence dependence by increasing the library of oligoribonucleotides used for RNA duplex melting experiments[36, 37], 2) experimentally determining energy parameters for small hairpin and internal loops[37, 42, 77], and 3) incorporating RNA sequence structure motifs determined through other avenues such as RNA comparative analysis through the

application of free energy “bonuses”[36]. The dynamic programming algorithm was improved to predict a collection of suboptimal structure structures for a given RNA sequence in addition to the minimum energy free energy structure[38]. By 1999, it was asserted that Mfold had an average prediction accuracy of 73% for RNA sequences up to 700 nt in length, and could reach as high as 97.1% if one considered a population of suboptimal secondary structure predictions in addition to the minimum free energy prediction[44].

However, in 1996, the accuracy of Mfold was measured using large collections of comparatively predicted 16S and 23S Ribosomal RNA (rRNA) secondary structure models which spanned the entire Tree of Life[40, 41]. It was determined that the average prediction accuracy of Mfold was 46% for 16S rRNA and 44% for 23S rRNA. Based on the results of Mathews et al.[44], Mfold had improved significantly since the 1996. Since the basic premise of Mfold had not changed between 1996 and 1999, I decided to test the claim of Mathews et al. that Mfold had improved significantly. Furthermore, I wanted to examine the viability of the paradigm of predicting the complex secondary structure of an RNA based solely on: 1) the nearest-neighbor thermodynamic model and 2) a dynamic programming algorithm which only considers the formation of base pairs and helices independently without consideration for either the arrangement of the helices in the final secondary structure or the kinetics governing the pathway by which the secondary structure is formed. In this chapter, I conduct a comprehensive evaluation of Mfold 3.1 using the largest available set of RNA comparatively predicted secondary structure models. This set of 1,411 structure models included 5S, 16S and 23S rRNA and tRNA and is well-balanced between short and long RNA sequences.

2.B HISTORICAL PERSPECTIVE AND SCIENTIFIC BACKGROUND

The establishment of the concept that an RNA can form a secondary structure in-vivo as a collection of DNA-style helical elements which consisted of intramolecular interactions[1, 3], was a significant achievement. In 1962 DeVoe and Tinoco were the first to establish the concept that the largest factor in the energy stabilization of a helix came from sequence dependent stacking interactions between the base-pairs[28]. These discoveries opened a new avenue of promising research; the application of thermodynamic models of RNA secondary structure for predicting a given RNAs secondary structure from its sequence. In the following section, I briefly discuss the evolution of the thermodynamic model for RNA secondary structure and its application to RNA secondary structure prediction. Due to the extensive amount of research conducted, I am forced to leave out the contributions of many in an effort to focus on the most important developments.

2.B.1 Mfold: Using the Nearest Neighbor Model and a Dynamic Programming Algorithm to Predict the Secondary Structure for a given RNA Sequence.

In 1971, Tinoco et al. proposed the first thermodynamic model for the free energy of forming secondary structure relative to the single strand[27]. This model was based on the concept that an RNA secondary structure helix is primarily stabilized by stacking interactions between the base pairs[28]. The free energy of the helix was sum of the free energies for forming consecutive A:U and G:C base pairs (e.g., 5' AU/AU 3') plus the destabilizing free energies of each hairpin, bulge or internal loop formed[27]. The destabilizing free energy of a loop relative to single strand was based on polymer theory for self-avoiding conformations and was a function of loop size[27]. This work established the nearest neighbor model for helix stability as the $\Delta G_{initiation} + \Delta G_{propagation}$ where $\Delta G_{propagation}$ is sum of the free energies for forming each base pair duplex in the

helix, and $\Delta G_{initiation}$ accounts for the destabilizing free energy of loop formation associated with forming the first base pair in the helix. In 1974, Borer et al. used melting experiments on 19 different RNA oligoribonucleotides to measure the effect of sequence dependence on helix stability[29, 31]. As a result, the thermodynamic parameters for 10 possible Watson-Crick nearest neighbor duplexes were experimentally determined[29, 31].

Subsequently, the first dynamic programming algorithms were introduced to predict the minimum free energy secondary structure for a given RNA sequence based on the experimentally determined thermodynamic parameters for RNA duplexes[32-34]. These algorithms were based on the assumption that an RNA secondary structure can be decomposed into a set of independent loops such that the nucleotides inside a loop can only interact with other nucleotides inside the loop. As a result of this assumption, it is possible to identify the minimum free energy secondary structure for a given RNA sequence by first identifying all the minimum energy base pairs and helices for a sequence, and then employing a traceback algorithm to construct the single minimum energy structure. It must be noted that one side effect of loop decomposition is that pseudoknotted structures can not be predicted. In 1989, Michael Zuker extended the dynamic programming algorithm to predict a series of suboptimal structures in addition to the minimum free energy structure for a given RNA sequence[38]. The collection of suboptimal foldings was intended to be a sampling of the possible foldings around the minimum energy folding rather than an exhaustive set.

In 1986, Turner et al. published a revised set of thermodynamic parameters for sequence dependence of RNA helix formation based on experiments from 45 RNA oligoribonucleotide sequences[35]. The empirical values for loop destabilizing energies of hairpin, bulge and internal loops of up to 30 nucleotides were tabulated; but lacked

significant experimental support. These revised thermodynamic parameters were combined with the dynamic programming algorithm of Zuker[33, 34] and released to the scientific community as Mfold 1.0 in 1987.

Between 1989 and 1994, improvements were made in the thermodynamic parameters. Destabilizing hairpin loop energies were modified based on comparative sequence analysis results that indicated a bias towards tetraloops (hairpin loops of size 4) with specific sequences[36, 78-80]. This was the first significant example of the application of comparatively predicted constraints in the prediction of RNA secondary structure via free energy minimization. Furthermore, the free energies of hairpin, bulge and internal loops (up to size 10) were determined based on experimental data including sequence dependence[36]. The sequence dependence was assumed to be a result of stacking interactions of terminal unpaired nucleotides on the adjacent base pair. For loops larger than 10 nt, Jacobson-Stockmeyer theory was used to extrapolate the length dependence on the free energy of loops[36]. In 1994, the thermodynamic effects of coaxial stacking of helices in RNA secondary structure were measured[37]. The measured effects were not included directly in Mfold. A separate program named *efn2* was created. The program *efn2* takes the set of optimal + suboptimal structure predictions for a given RNA sequence and re-computes the free energies of all structure predictions based on observed coaxial stacking interactions. These improvements were included in Mfold version 2.3 which was made available to the scientific community in 1995[37, 38].

2.B.2 Evaluating the Accuracy of Mfold 2.3 using a set of Comparatively Predicted 16S and 23S rRNA secondary structures

Using a set of 15 comparatively predicted 16S rRNA sequences, the average accuracy of Mfold 2.3 was reported to be 49%[81]. Subsequently, in 1995 and 1996 the Gutell Lab published two studies quantifying the accuracy of Mfold 2.3[37, 38] using the

largest sets available at the time of comparatively predicted 16S and 23S rRNA sequences[40, 41]. The individual 16S and 23S rRNA sequences were selected from all major branches of the Tree of Life; Archaea, Bacterial, and Eukaryotic[40, 41]. Furthermore, for the Eukaryotes, they included Chloroplast and Mitochondrial encoded rRNA sequences in addition to Nuclear rRNA sequences[40, 41]. The purpose of sampling the 16S and 23S rRNAs in this manner was to attempt to quantify the accuracy of Mfold 2.3 while considering the wide variety of secondary structure models possible for 16S and 23S rRNA. The secondary structure conservation diagram for 16S rRNA in Figure 2.1 demonstrates of sequence and structural variation and conservation observed across the Tree of Life. The most important observations from these studies included: 1) the average prediction accuracy for Mfold 2.3 was 46% for 16S rRNA and 44% for 23S rRNA; 2) The prediction accuracy for a 16S or 23S rRNA could be as high as 80% and as low as 10%; 3) Pairings where the 5' and 3' nt were separated by no more than 100 nucleotides, termed "short-range", were predicted the most accurately; 4) Archael rRNAs were predicted with the highest accuracy followed in decreasing order by Bacterial, Chloroplast, Mitochondrial and Eukaryotic Nuclear[40, 41].

2.B.3 Mfold 3.1: Improvements in the Sequence Dependence of Thermodynamic Parameters and Multistem Loop Prediction

In 1999, Mathews et al. published a paper introducing Mfold 3.1[44]. In their paper, Mathews et al. describe a number of improvements to thermodynamic model used within Mfold through the addition of new sequence-dependent energetic measurements for the Watson-Crick paired helices, small internal loops, hairpin loops, multibranch (i.e., multistem loops) loop initiation and coaxial stacking[42, 43]. Thermodynamic parameters for multistem loop initiation based on tuning Mfold with a set of known RNA secondary structure models [44].

Mathews et al. evaluated the accuracy of Mfold 3.1 using a set of 955 comparatively predicted RNA secondary structures covering 8 different RNAs. They reported that on average, the predicted minimum energy secondary structure contained 73% of the comparatively predicted Watson-Crick and G:U base pairs when considering sequences of up to 700 nucleotides[44]. Furthermore, they reported that this number could reach as high as 86% average accuracy within a single predicted secondary structure model if one considers a collection of 750 possible suboptimal predictions[44]; a collection which on average contains 97.1% of the expected base pairs[44].

With their assessment, Mathews et al were able to demonstrate significant improvement in the accuracy of Mfold through: 1) the addition of more sequence dependent energy parameters to the thermodynamic model for RNA duplexes, small internal loops and hairpin loops and 2) tuning the multistem loop initiation parameters using a larger set of RNA comparative structure models. However, one must consider that their results included a number of caveats: 1) it was necessary to restrict the length of sequences to 700 nucleotides[44], 2) the suboptimal population of predicted secondary structures was allowed to have as much as 20% difference in free energy compared to the optimal structure prediction[44], and 3) their comparative RNA structure population was skewed towards the shorter RNAs, tRNA and 5S rRNA.

2.C RE-EVALUATING THE ACCURACY OF MFOLD 3.1

The assessment of the accuracy of Mfold 3.1 by Mathews et al.[44] was provided with important caveats as I mentioned in Section 2.B.3. Using the comprehensive analysis of Mfold 2.3[37, 38] conducted in 1995[40] and 1996 [41], (Section 2.B.2) as a guide; in 2003 I conducted the largest and most detailed evaluation of the accuracy of Mfold 3.1 using an RNA secondary structure database from the CRW Project[62] which consisted of 1,411 comparatively predicted structure models for rRNA and tRNA

spanning the three major phylogenetic domains. I analyzed over 1.5 million nucleotides and more than 380,000 comparatively predicted base pairs. The accuracy of the comparatively predicted rRNA secondary structure models has been established using high-resolution crystal structures for the 30S and 50S ribosomal subunits[25] (Section 3.B.4). The most recent comparative models for 16S and 23S rRNA are based on alignments of 7000 and 1050 sequences respectively; each alignment spanning the entire Tree Of Life[62].

The primary goal of this evaluation was to determine if Mfold 3.1 was significantly more accurate than Mfold 2.3 using a large RNA comparative structure database; well-balanced between shorter and longer RNAs. *Furthermore, this evaluation will assess the population of suboptimal structure models predicted for any given RNA sequences, and determine their potential contribution to improving the accuracy of Mfold. The most significant results of this evaluation will be: 1) a better characterization of how well the additional constraints to the energy parameters and the secondary structure prediction algorithm improved the prediction accuracy of Mfold and 2) to identify remaining weaknesses which will require additional constraints and enhancement of the thermodynamic model or the secondary structure prediction algorithm.*

To establish the feasibility of comparing the results of this evaluation with results from previous Gutell Lab studies[40, 41] one must determine the extent to which the comparatively predicted structure models have been revised between 1995 and 2003. Only minor differences exist between the 1995 and 2003 versions of the 16S and 23S rRNA comparative structure models. For example, in the 2003 version of the *Haloferax volcanii* 16S rRNA secondary structure model[62], 30 base-pairs were added, 17 base-pairs were removed, and 427 base-pairs remained unchanged. The result of the

modifications is a net difference of approximately 3% in the total number of base-pairs in the model when compared to the *Haloferax volcanii* 16S rRNA secondary structure model used in the 1995 Gutell Lab study[40]. Similar numbers were observed for the other comparatively predicted structures evaluated.

The complete evaluation was published in BMC Bioinformatics in 2004[1][82], and the results and discussion in the following sections were published in that paper. A detailed web site containing results of the study is available at the CRW Web Site[62] and selected results sets from that web site are included in Appendices A-D.

2.C.1 RNA Comparative Structure Database

The comparative structure database assembled for this evaluation was the largest ever used for comparing RNA secondary structure models predicted by comparative analysis and the Mfold program. The 1995 and 1996 evaluations conducted by the Gutell Lab analyzed only 56[40] and 72[41] RNA sequences respectively. The 1999 study by Mathews et al. analyzed a total of 151,503 nucleotides and 43,519 comparatively predicted canonical base-pairs (i.e., G:C, A:U and G:U) from 955 sequences[44].

As shown in Table 2.1, the evaluation encompassed a total of 1,411 RNA sequences, encompassing 1,505,143 nucleotides and 385,854 canonical secondary structure base-pairs. Sequences from each of the three major phylogenetic domains, the Archaea, Bacteria, and Eucarya were included. The Eukaryotic segment of the database included sequences encoded in the Nucleus, Chloroplast and Mitochondrion. Table 2.1 depicts the distribution of RNA secondary structures included in evaluation as a function of phylogenetic classification. Of the 1,411 sequences analyzed, 569 were tRNA, 496 were 16S rRNA, 256 were 23S rRNA, and 90 were 5S rRNA. 47% of the RNA

[1] The author's contributions are discussed at the end of the paper: 82. Doshi, K.J., et al., *Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction.* BMC Bioinformatics, 2004. **5**(1): p. 105..

sequences in our dataset contained 700 or less nucleotides while 53% had more than 700 nucleotides. Detailed statistics for all 1,411 RNA sequences included in the dataset are available in Appendix A.

By comparison, the set of RNA comparative structure models selected by Mathews et al. contained 8 different RNAs (5S, 16S and 23S rRNA, Group I and II introns, RNaseP, SRP and tRNA)[44]; however, 86% of the 955 structure models were rRNA (5S, 16S and 23S) and tRNA, and 96% of these sequences had less than 700 nucleotides[44]. The supplementary material provided by Mathews et al. does not report the lengths of all 5S rRNA and tRNA sequences used. For estimation purposes, I assume that all 5S rRNA and tRNA sequences used by Mathews et al. were less than 700 nucleotides. The 22 16S and 5 23S rRNA sequences selected included sequences from each of the three major phylogenetic domains[44].

The sequence diversity of the RNA comparative structure database used in this evaluation was quantified by calculating sequence identity for all pairs of 16S and 23S rRNA sequences within the different phylogenetic classifications. For the 16S rRNA dataset, 75% of the Archaeal, 86% of the Bacteria, 71% of the Chloroplast, 99% of the Mitochondrial, and 94% of the Eukaryotic Nuclear sequence pairs had less than 80% sequence identity, while only 4% or fewer of the pairs in a given phylogenetic classification had 95% or more sequence identity (Figure 2.2). Moreover, 79% of the Mitochondrial and 48% of the Eukaryotic Nuclear 16S rRNA sequence pairs had less than 50% sequence identity (Figure 2.2). The 23S rRNA dataset exhibited even more diversity than the 16S rRNA dataset, as 87% of the Archaeal, 94% of the Bacteria, 76% of the Chloroplast, 99% of the Mitochondrial, and 97% of the Eukaryotic Nuclear sequence pairs had less than 80% sequence identity, while 2% or fewer of the sequence pairs in a given phylogenetic classification had more than 95% sequence identity (Figure

2.2). The detailed results of the pairwise sequence identity calculations are available in Appendix B.

The comparative structure database used by Mathews et al. utilized known modified nucleotide information in tRNA to limit the base-pairing potential for those nucleotides that are modified[44]. In my evaluation, rRNA or tRNA base modifications were not taken into account. A simple analysis of the tRNA dataset from my evaluation shows that 70% of the tRNA sequences came from genomic DNA sequences: as a result, no modification data was available for those sequences. For the remaining 30%, the number of modifications that could prevent A-form helix formation was minimal; between only 1 to 5 modifications per sequence.

2.C.2 Computational Setup

Evaluating the accuracy of folding 1,411 RNA sequences was not a computationally easy task in early 2002. To facilitate folding the RNA sequences and storing and cataloguing the results, I developed a distributed, workflow based job management system which utilized a relational database and was capable of automatically managing the entire computational process. The idea for this workflow system was based on a distributed, object-oriented database schema evolution tool I developed while working at UOP LLC. in 1999. Figure 2.3 is schematic diagram of the system. The folding computation was broken up into four specific tasks, which were required to be performed in order for each sequence: 1) preparing the RNA sequence for folding with Mfold 3.1, 2) folding the sequence with Mfold 3.1, 3) compressing the results into an archive, and 4) extracting the results from the archive and inserting them into a separate results database for subsequent analysis. The four specific tasks were entered into the "Task Table" in the relational database for each of the 1,411 RNA sequences (Figure 2.3) and assigned one of four different states: HOLDING, READY,

PROCESSING, and FINISHED. Since the four tasks had to be processed in order for each RNA sequence, only the first task was marked READY and the subsequent tasks were marked HOLDING. Dependencies between tasks were entered into the database such that when a given task was completed, dependent tasks which were in the HOLDING state could be updated to the READY state. All task specific arguments such as the name of the sequence or the name of the output file, etc. were placed in "Task Specific Tables" (Figure 2.3).

A dispatcher server and a task processing client were implemented in Java. The dispatcher server was responsible for querying the relational database to identify all tasks currently in the READY state. The task processing clients were responsible for executing a given task. The interaction between the dispatcher server and task processing clients was straightforward. When a given task processing client was available, it would request a new task from the dispatcher server (Figure 2.3, Step 1). The dispatcher server would identify a task that was in the READY state (Figure 2.3, Step 2), update the state of that task to PROCESSING (Figure2.3, Step 3) and assign that task to the processing client (Figure 2.3, Step 4). Upon completion of the task, the task processing client would update the state of that task in the relational database to FINISHED (Figure 2.3, Step 5). Furthermore, the state of any dependent task(s) was updated from HOLDING to READY (Figure 2.3, Step 5). The dispatcher server and task processing clients were all started on different computers, effectively creating a distributed, parallel processing environment (Figure 2.3). All 1,411 RNA sequences in the comparative structure database were processed in under 48 hours using my workflow-based distributed job management system. After compression, the aggregate set of folding results required over 150 GB of disk space (which was still considered large in 2002). The results database contained over

100 tables and were retrieved on demand, using simple SQL queries, when required to calculate final results.

2.C.3 Mfold Folding Parameters

The most important parameters used to control RNA secondary structure prediction by Mfold are window size (W), percent suboptimality (P), and the inclusion or exclusion of additional energy calculations based on coaxial stacking and multistem loop scoring (*efn2*)[37, 44]. The percent suboptimality establishes the energy range for computed foldings. The range is ΔG_{min} to $\Delta G_{min} + \Delta\Delta G$, where $\Delta\Delta G$ is a percentage of ΔG_{min} [38]. The window size establishes the amount of difference between the suboptimal folds by requiring that given folding has a minimum number of base pairs that are a minimum distance from any base pairs in suboptimal folds already computed. For a given Mfold invocation, the program *efn2* is used to re-compute the energetics of each predicted structure based on coaxial stacking and multistem loop opportunities. The predicted structures are then re-ordered by the modified ΔG and a new optimal structure is selected. The reader should note that coaxial stacking is only considered after the dynamic programming algorithm has selected the possible base pairs. In contrast, multistem loop energetics are considered during the execution of the dynamic programming algorithm and with the *efn2* program in the subsequent step. However, the energy function used within the dynamic programming grows for multistem loops is very simplistic, the free energy increases linearly with the length of the loop[44]. In the *efn2* the linear energy function for multistem loops is replaced with a function that grows logarithmically with the length of the loop.

Previous evaluations by the Gutell Lab[40, 41] used window sizes (W) of 10 and 20, respectively, with no *efn2* re-evaluation; the selection of window size was limited by the computational resources available at the time the studies were conducted. Mathews et

al.[44] evaluated Mfold 3.1 with a window size (W) of 0, percent suboptimality (P) of 20%, and *efn2* re-evaluation. Each of the 1,411 sequences in this evaluation was folded, using a window size (W) of 1, percent suboptimality (P) of 5%, and maximum number of suboptimal foldings (MAX) of 750. The optimal, or minimum, free energy prediction and 749 suboptimal predictions were determined after re-ordering the structure predictions by the efn2 re-computed energetics. The reader should note that some sequences did not yield 749 suboptimal structure predictions with the folding parameters specified. The folding parameters used in this evaluation were selected to: 1) maximize the number of structures predicted for any given sequence, 2) densely sample the suboptimal population close to the minimum free energy structure, and 3) to include coaxial stacking in the energy calculations with the *efn2* option in Mfold 3.1.

The previous Gutell Lab evaluations used a significantly smaller window size. Since the difference in window size affects the number and diversity of suboptimal structures computed, and the previous Gutell Lab studies did not include any energy re-computation and re-ordering of predicted structures with *efn2*, this difference will not have a significant impact on the results. The evaluation by Mathews et al. used different values for percent suboptimality (P) and window size (W) in computing suboptimal structure predictions. The net result of the difference is that the Mathews et al. study considered suboptimal structures with energy values further away from the minimum free energy prediction than my evaluation. This difference could have an impact on the results since the Mathews et al. study may include a structure prediction for a given sequence that is extremely unfavorable energetically (and would be excluded from the suboptimal population of this evaluation) but upon *efn2* re-ordering becomes the minimum energy structure. However, if this scenario is realized one could consider that to be strong

evidence that the basic energy function in Mfold does not adequately model RNA secondary structure.

2.C.4 Results: RNA Secondary Structure Prediction Accuracy for Mfold 3.1

The compilation of the secondary structure prediction accuracies for each of the 1,411 RNA sequence in this evaluation are presented in detail in Appendix C. In this section I present results of the evaluation including: 1) the raw secondary structure prediction accuracies (Section 2.C.4.1), 2) the statistical variation in secondary structure prediction accuracy of the Mfold minimum energy structure prediction (Section 2.C.4.2), 3) the variation in secondary structure prediction accuracy as a function of phylogenetic classification (Section 2.C.4.3) and 4) a direct comparison between Mfold 2.3 and Mfold 3.1 for selected RNA comparative structure models (Section 2.C.4.4). See Section 2.G.1 for a discussion on how prediction accuracies were computed.

2.C.4.1 Raw Prediction Accuracy

The overall average prediction accuracy for the 1,411 RNA sequences in this evaluation was 54% (Appendix C) compared to 73% reported by Mathews et al.[44]. The average folding accuracies per sequence for 5S rRNA and tRNA, the two smallest molecules in my evaluation, were 71% and 69% respectively (Table 2.2). The evaluation by Mathews et al. reported average accuracy per sequence of 78% for 5S rRNA and 83% for tRNA[44]. Accuracies for the sets of 5S rRNAs and tRNAs in my evaluation were about 25% higher than the average accuracies for the 16S (41%) and 23S (41%) rRNAs (Table 2.2). By comparison, the Gutell Lab's previous evaluations of Mfold 2.3 reported an average folding accuracy of 46% for 16S rRNA and 44% for 23S rRNA[40, 41], and the evaluation by Mathews et al. of Mfold 3.1 reported average accuracies (for folding complete RNA sequences) of 51% for 16S rRNAs and 57% for 23S rRNAs[44]. When

considering only sequences analyzed in previous Gutell Lab studies, the average prediction accuracy of Mfold 3.1 was 45% for 16S rRNA and 43% for 23S rRNA (Table 2.2).

2.C.4.2 Variation in Prediction Accuracy

To gauge the variation in accuracy score for the optimal structures predicted with Mfold 3.1, the percentages of scores greater than 60% and less than 20%, the median accuracy score, and the highest and lowest accuracy scores were identified for the four RNA types in my evaluation (Table 2.2). This analysis revealed a large range of accuracy scores with values significantly larger and smaller than the respective average value. The highest accuracy score for the optimal structure for each RNA type was 100% for tRNA (i.e., at least one of the predicted tRNA structures had 100% of the base-pairs in the comparative model), 98% for 5S rRNA, 77% for 16S rRNA, and 74% for 23S rRNA (Table 2.2). In contrast, at least one of the optimal folds for 5S rRNA or tRNA had a score of 0 (i.e., none of the base-pairs in the comparative structure model were predicted with Mfold). The lowest accuracy score was 5% for 16S rRNA and 1% for 23S rRNA (Table 2.2).

The median accuracy score observed for each RNA type was 70% for tRNA, 81% for 5S rRNA, 41% for 16S rRNA and 41% for 23S rRNA (Table 2.2). For 16S and 23S rRNA the overwhelming majority (86% for 16S rRNA and 89% for 23S rRNA) of optimal structures predicted had an accuracy score greater than 20% and less than 60% (Table 2.2), a trend also observed in previous Gutell Lab evaluations[40, 41]. The majority of optimal structures predicted for 5S rRNA (77%) had an accuracy score greater than 60% (Table 2.2). For the tRNA, 60% of the optimal structures were predicted with accuracy greater than 60% and 39% of the optimal structures predicted with accuracy between 20% and 60%. The percentage of predicted structures with an

accuracy score below 20% was highest for 23S rRNA (6%), followed by 16S rRNA (4%), 5S rRNA (4%), and tRNA (2%) (Table 2.2). A complete list of accuracy scores for all 1,411 RNA sequences in the evaluation dataset are available in Appendix C.

2.C.4.3 Prediction Accuracy as a Function of Phylogeny

The previous Gutell Lab evaluations revealed significant variation in the prediction accuracy scores of Mfold 2.3 within and between the five major phylogenetic groups[40, 41]. For the 16S rRNA dataset in my evaluation, the Archaeal sequences had the highest average accuracy (62%), while the Mitochondrial sequences had the lowest average accuracy (30%). Between these two extremes were the Bacteria (49%), Chloroplast (46%), and Eukaryotic Nuclear (34%) sequences (Table 2.3). These results were consistent with previous with results [40], except that the Archaeal and Bacterial accuracy scores were slightly lower in my evaluation (62% vs. 68% and 49% vs. 56%). For 23S rRNA, the Archaeal dataset again had the highest accuracy scores (58%), followed by the Bacterial (49%), Eukaryotic Nuclear (42%), Chloroplast (39%), and Mitochondrion (30%) (Table 2.3). These results were also consistent with the trends observed previously[41].

2.C.4.4 Direct Comparison Between Mfold 2.3 and Mfold 3.1 for Selected RNA Comparative Structure Models

To access specific differences between the optimal foldings from the evaluation of Mfold 2.3[40, 41] and my evaluation of Mfold 3.1 for select 16S and 23S rRNA sequences, base pairs predicted with both versions of Mfold were mapped onto the comparative structure models for each sequence. Some of the base-parings were predicted correctly with both versions of Mfold, other base pairings were predicted exclusively by one version, while a third set of base pairings were not predicted correctly with either version. The *Haloferax volcanii* 16S rRNA (Figure 2.4) (A) and

Thermococcus celer 23S rRNA (Figure 2.4) (B.1, B.2) sequences were generally predicted very well with both versions of Mfold (77% and 67% accuracy in the current evaluation compared with 81% and 74% previously[40, 41]. Meanwhile, *Giardia intestinalis* 16S (Figure 2.4) (C) and 23S (Figure 2.4) (D.1, D.2) rRNA sequences were predicted poorly with both versions of Mfold (23% and 33% accuracy in the current evaluation compared with 10% and 24% previously[40, 41]. The base pairings in the comparative model that were missed by both versions of Mfold were generally longer range (Section 2.C.5). The poor prediction accuracy for *G. intestinalis* 16S and 23S rRNA with both versions of Mfold (Figure 2.4) (C, D.1, D.2) was representative of other sequences originally predicted with low accuracy by Mfold 2.3. A total of 9 out of 10 16S sequences and 7 of the 8 23S sequences predicted with accuracy of 30% or less with Mfold 2.3[40, 41] were still predicted with less than 30% accuracy with Mfold 3.1 (Table 2.4).

2.C.5 Results: RNA Secondary Structure Prediction Accuracy and the RNA Contact Distance

For a given protein, the average sequence separation between pairs of amino acids involved in non-covalent interactions is defined as the "Contact Order"[83]. Two similar topological descriptions for non-covalent interactions in RNA are: 1) "RNA Contact Distance," the number of intervening nucleotides between two nucleotides that base pair and 2) "RNA Contact Order," the average of the RNA Contact Distances for a given RNA sequence. We considered any base pair with an RNA Contact Distance of 100 or less to be "*short-range*," a RNA Contact Distance of 101-501 to be "*mid-range*," and a RNA Contact Distance of 501 or greater to be "*long-range*." Figure 2.5 is a graphical depiction of the RNA Contact Distance for all base pairs in the 16S rRNA comparative structure model. The majority of base-pairs in the 16S and 23S rRNA secondary structure

models predicted with comparative analysis were short-range (Table 2.5), and the previous Gutell Lab evaluations of Mfold 2.3 had established that short-range base pairs are predicted more accurately than long-range base pairs[40, 41] (Section 2.B.2). My evaluation examined the effects of RNA Contact Distance on prediction accuracy by comparing: 1) the accuracies of the short-range and long-range base pairs predicted with Mfold 3.1 and Mfold 2.3 (Section 2.C.5.1), 2) the distribution of short-, mid-, and long-range base pairs in the comparative models predicted by Mfold 3.1 (Section 2.C.5.2), and established a relationship between the base-pair prediction accuracy and the contact distance for 16S rRNA (Section 2.C.5.3).

2.C.5.1 Prediction Accuracy for Short-Range and Long-Range Base Pairs

The 496 16S rRNA comparative structure models in my evaluation were comprised of 191,994 canonical base-pairs. A total of 145,058 (76%) of these base pairs were short-range. 75,763 (52%) of these base pairs were predicted correctly (Table 2.5), and the average accuracy for short-range base pairs was 50% per sequence (Table 2.5). By comparison, an average accuracy of approximately 55% per sequence was observed previously for short-range base pairs[40](Table 2.5). A total 3,932 (2%) of the comparatively predicted base pairs in the set of 496 16S rRNA comparative structure models evaluated were long-range, and only 193 (5%) of these base pairs were predicted correctly (Table 2.5). If one considers all 46,936 (24%) comparatively predicted base-pairs with an RNA Contact Distance greater than 100, a significantly lower percentage are predicted correctly 6,171 (13%) when compared with the prediction accuracy for base pairs with an RNA Contact Distance less than 100 (52%) (Table 2.5)

In order to provide an accurate and complete analysis, I must address the long-range pseudoknotted base-pairs in 16S rRNA. Pseudoknotted base pairs were scored in this evaluation; however, Mfold is not capable of predicted pseudoknotted base pairs due

to the implementation of the dynamic programming algorithm (Section 2.G.1). Three long-range base-pairs (17:918, 18:917, and 19:916, Escherichia coli numbering) are found in over 99% of the 16S rRNAs in the three phylogenetic domain two organelle alignment[62] (Figure 2.6). To determine the extent by which pseudoknotted base pairs are influencing the prediction accuracy of long range comparative base pairs by Mfold 3.1, I can estimate the effect of removing these three pseudoknotted base pairs from the evaluation. If I make the conservative assumption that each of the 496 16S rRNA comparative structure models in my evaluation has these three pseudoknotted base pairs, the total number of long-range comparatively predicted base pairs expected without considering pseudoknotted base pairs is reduced from 3,932 (Table 2.5) to 2,544 The percentage of long-range base-pairs predicted correctly would increase from 5% (193/3,932) to 8% (193/2544). Including these long-range pseudoknotted base pairs has minimal impact on the results.

The 256 23S rRNA comparative structures models in my evaluation contained a total of 178,958 canonical base-pairs. 134,085 (75%) of these canonical base pairs were short-range, and 67,130 (50%) of those base pairs were predicted correctly (Table 2.5), and the average prediction accuracy for short-range base pairs was 47% per sequence (Table 2.5). An average accuracy of approximately 53% per sequence was observed previously[41] (Table 2.5). A total of 7,752 (4%) of the comparatively predicted base pairs in the set of 256 23S rRNA comparative structure models evaluated were long-range, and 1,317 (17%) were predicted correctly (Table 2.5).

2.C.5.2 Distribution of Predicted Base Pairs by RNA Contact Distance

A total of 223,957 base-pairs were predicted with Mfold 3.1 (optimal structure predictions only) for the 496 16S rRNA comparative structure models in my evaluation (Table 2.6). This was 31,963 more than expected. Of the 223,957 base-pairs, 150,886

(67%) were short-range, 43,498 (19%) were mid-range and 29,573 (13%) were long-range (Table 2.6). Of the 150,886 short-range base-pairs, 75,763 (50%) were correct while only 193 (0.7%) of the long-range base-pairs were correct (Table 2.6). 13% of the total number of 16S rRNA base-pairs predicted with Mfold 3.1 were long-range while only 2% of the comparatively predicted base-pairs were long-range.

Similar results were observed for the 256 23S rRNA comparative structure models in my evaluation. A total of 218,908 base-pairs were predicted by Mfold 3.1 (optimal structure predictions only). This was 39,950 more than expected. Of the 218,908 base-pairs, 137,780 (63%) were short-range, 44,139 (20%) were mid-range and 36,989 (17%) were long-range (Table 2.6). Of the 137,780 short-range base pairs, 67,130 (49%) were correct, while only 1,317 (4%) of the long-range base pairs were correct (Table 2.6). Akin to the 16S rRNA dataset, 17% of the total number of 23S rRNA base pairs predicted with Mfold 3.1 were long-range while only 4% of the comparatively predicted base-pairs were long-range.

2.C.5.3 Relationship Between Prediction Accuracy and RNA Contact Distance

The results reported in Sections 2.C.5.1 and 2.C.5.2 prompted a more sophisticated analysis to quantify the relationship between the accuracy of base-pairs predicted with Mfold 3.1 and RNA Contact Distance. Figure 2.7 shows the distribution of contact distances for the 191,994 canonical base pairs from the 496 16S rRNA comparative structure models in this evaluation. The frequency of comparatively predicted base pairs is observed to decrease exponentially as RNA Contact Distance increases (Figure 2.7). Based on this observation, I divided the 191,994 16S rRNA comparative base-pairs into seven somewhat equally-sized bins (within one order of magnitude of one another) by considering the contact distance values on a logarithmic scale instead of a linear scale (Section 2.G.3).

The Mfold 3.1 average prediction accuracy for base pairs in each of the seven bins was determined. The prediction accuracy was 61% for base pairs in the smallest contact distance bin, 3-8, 57% for base pairs in the 9-19 contact distance bin, 47% for base pairs in the 20-47 bin, 46% for the 48-117 bin, 15% for the 118-293 bin, 7% for the 294-733 bin, and 0% for the 734-1833 bin (Figure 2.8). The approximately linear relationship obtained from plotting the accuracy for logarithmically-scaled bins confirms an exponential relationship between the accuracy of Mfold and the RNA Contact Distance (Figure 2.8).

2.C.6 Results: RNA Secondary Structure Prediction Accuracy including the Suboptimal Population for 16S rRNA

The initial dynamic programming algorithms for RNA Secondary Structure prediction via free energy minimization only returned the optimal or minimum energy structure prediction[32-34]. Subsequently, Michael Zuker extended the dynamic programming algorithm to compute a set of suboptimal structure predictions within a specified $\Delta\Delta G$ of the minimum free energy structure prediction[38]. However, Zuker's algorithm does not perform an exhaustive analysis of the suboptimal population, but rather a sampling. In their 1999 paper, Mathews et al. used the suboptimal population of predicted structure models for a given RNA with had a free energy within 20% of the minimum free energy predicted structure as an example of the extent to which Mfold 3.1 was capable of accurately predicting RNA secondary structure[44]. They found that an average of 77% of the comparatively predicted base pairs for a given 16S rRNA could be identified in the suboptimal population (when folding complete RNA sequences) [44]. However, by using such a large energy range, Mathews et al. are heavily discounting the extent to which the energetic parameters are capable of distinguishing the correct RNA secondary structure from the astronomical number of possible secondary structures.

In this section, the population of suboptimal predictions was analyzed for each of the 496 16S rRNA comparative structures models in my evaluation. The population of suboptimal predictions for any given 16S rRNA in my evaluation was constrained such that each suboptimal prediction was required to have a free energy within 5% (Section 2.C.3) of the optimal structure prediction. Because of this constraint, the suboptimal population should provide a better measure of the extent to which the energetic parameters are capable of identifying the 16S RNA secondary structure for a given sequence. In Section 2.C.6.1, the set of unique comparative base pairs present is identified from the collection of all base-pairs predicted in the suboptimal population for each individual 16S rRNA sequence. In Section 2.C.6.2, the distribution of comparative base pairs identified in Section 2.C.6.2 as a function of RNA Contact Distance is determined.

2.C.6.1 Unique Comparatively Predicted Base Pairs Identified in the Suboptimal Population

The 496 16S rRNA comparative structure models in this evaluation contained a total of 191,994 unique canonical, comparative base pairs (Table 2.7). 81,934 of these canonical base pairs were predicted with Mfold 3.1 to be in a minimum free energy structure prediction (Table 2.7), with an average accuracy of 41% per sequence (Table 2.7) (Section 2.C.4.1). When considering the entire suboptimal population of structure predictions for each of the 496 16S rRNA sequences, a total of 137,000 comparative canonical base pairs were predicted correctly by Mfold, and the average accuracy per sequence increased to 71% (Table 2.7). This represented a 30% increase in the average number of base pairs in the comparative model that were predicted correctly per sequence. The average accuracy per sequence for an Archaeal, Bacterial, Eukaryotic Nuclear, Chloroplast, and Mitochondrial sequence increased by 21%-41% respectively

(Table 2.7), and the largest increase for a single sequence (68%) was observed in the Mitochondrial dataset (Table 2.7). Of course these dramatic improvements in accuracy were offset by a significant increase in the number of base pairs predicted incorrectly; Mfold experienced a large drop in selectivity. The total number of unique incorrect base pairs predicted for the 496 minimum free energy structure predictions was only 142,023, while the total number of unique incorrect predictions was 2,372,305 for the 496 sets of optimal plus 749 suboptimal structure predictions, a 1,664% increase in the number of incorrect predictions (Table 2.7).

Section 2.C.6.2 Distribution of Comparatively Predicted Base pairs Identified in the Suboptimal Population as a Function of RNA Contact Distance

As noted in Section 2.C.6.1, a total of 137,000 canonical, comparatively predicted base pairs are identified in the suboptimal populations from each of the 496 16S rRNA comparative structure models in this evaluation. The distribution of these comparatively predicted base pairs identified in the suboptimal population as a function of RNA Contact Distance is displayed in Table 2.8. 76% of the base pairs were short-range (RNA Contact Distance less than 101), 22% were mid-range (RNA Contact Distance of 101-500), and 2% were long-range (RNA Contact Distance greater than 500). 54,994 (29%) of the total canonical, comparative base pairs from the 496 16S rRNA comparative structure models in this evaluation are never identified when one considers the suboptimal populations (Table 2.8); an average of 111 base pairs per 16S rRNA comparative structure model. 29,587 (54%) of the base-pairs were short-range, 22,935 (42%) were mid-range, and 2,472 (4%) were long-range (Table 2.8). If one considers just the long-range base pairs, the 496 16S rRNA comparative structure models contain 3,932 such base pairs (2% of the total), and as noted in Section 2.C.5.2, Mfold predicts only 193 long range base pairs correctly (which is 0.7% of the 29,573 total long range base-pairs predicted) when one

only considers the minimum free energy structure prediction (Table 2.6). If one includes the suboptimal populations for the 496 16S rRNA comparative structure models, a total of 1,460 long range base pairs are predicted correctly (Table 2.8); an increase of 656%. Subsequently, 2,472 comparatively predicted long-range base pairs are still not identified; 63% of the total (Table 2.8). If one excludes the long-range pseudoknotted base pairs in 16S rRNA as discussed in Section 2.C.5.1, the number of long-range comparative base pairs never identified is reduced to 1,084 from 2,472, which is only 43% of total comparatively predicted long-range base pairs that can be identified by Mfold 3.1.

2.D IMPORTANT CONCLUSIONS FROM THE EVALUATION OF MFOLD 3.1

As discussed in Section 2.B, the nearest-neighbor thermodynamic model that governs RNA secondary structure was originally postulated in 1971[27], and the first sequence dependent energy parameters derived from melting experiments on short oligoribonucleotides was presented in 1974[29-31]. In the intervening time period between 1974 and 1999, a significant amount of research was conducted to enhance the thermodynamic model by conducting experiments to refine parameter estimation (Section 2.B.1, 2.B.3). The expectation was that the improved thermodynamic model would reflect the increasing knowledge about the complexity of RNA structure as determined through methods such as comparative analysis. In 1999, Mathews et al. claimed that significant refinements had been made in the sequence dependence of the energetic parameters of the thermodynamic model underlying Mfold 3.1; implying that it was sufficiently rigorous such that it could on average predict up to 73% of known base-pairs, given a number of caveats (Section 2.B.3)[44]. The evaluation of Mfold 3.1 discussed in this chapter was designed to determine 1) if Mfold 3.1 was significantly more accurate than Mfold 2.3 using a large RNA comparative structure database; well-balanced

between shorter and longer RNAs and 2) in what areas could Mfold 3.1 be improved? The most important conclusions from the evaluation were:

1. While the prediction accuracies for shorter RNAs such as 5S rRNA and tRNA are in agreement between Mathews et al.[44] and this evaluation, overall Mfold 3.1 was not more accurate than Mfold 2.3 when considering an RNA comparative structure database that is better balanced between shorter RNAs and longer RNAs. This conclusion is corroborated by multiple results from the evaluation in Section 2.C.4. Prediction accuracies for 16S and 23S rRNA comparative structure models showed little improvement between Mfold 2.3 and Mfold 3.1 (45% and 42% for Mfold 2.3 vs. 41% and 41% for Mfold 3.1) (Section 2.C.4.1). Furthermore, many specific 16S and 23S rRNA sequences predicted poorly with Mfold 2.3 were still predicted poorly with Mfold 3.1 (Section 2.C.4.4), and the phylogenetic dependence of the prediction accuracy for 16S and 23S rRNA observed in the current evaluation was consistent with results from the previous Gutell Lab studies (Section 2.C.4.3). If one considers the increased sample size and the inclusion of pseduoknots (Section 2.C.4.5) in the accuracy scoring for the evaluation in Section 2.C, one can conclude the Mfold has not improved significantly for longer RNAs.

The study by Mathews et al. reports: 1) significantly higher overall optimal prediction accuracy for Mfold 3.1 (73% vs. 54% in the current evaluation) (Section 2.C.4.1); 2) average optimal prediction accuracies as high as 66% and 70% for 16S and 23S rRNA[44]; 3) significantly higher average optimal prediction accuracy for tRNA (83% vs. 69% in the current evaluation) (Section 2.C.4.1). However, one must consider the important differences between the Mathews et al. study and the evaluation in Section 2.C: 1) The sampling of RNA comparative structure models used by Mathews et al. was heavily weighted towards RNAs with less than 700 nucleotides (Section 2.C.1); 2) The

best optimal prediction accuracies for 16S and 23S rRNA reported by Mathews et al. were based on segmenting the sequences into smaller domains[44], and when Mathews et al. folded entire 16S and 23S rRNAs, their reported prediction accuracies were significantly lower (51% and 57%) (Section 2.C.4.1); 3) When scoring a base pair as predicted correctly, Mathews et al. allowed either the 5' or 3' half of the predicted base pair to be off by 1 nucleotide (Section 2.G.1); 4) Bases known to be modified in tRNA that are subsequently unable to fit into an A-form helix were constrained to be unpaired (Section 2.C).

2) The significant decrease in the accuracy of Mfold for longer RNA sequences is partially due to the inability of Mfold to predict base-pairs with large RNA Contact Distance ("long-range") accurately and reliably. This conclusion is corroborated by multiple results from the evaluation in Section 2.C.5. Previous Gutell Lab studies had established a correlation between the prediction accuracy of base pairs in 16S and 23S rRNA and their RNA Contact Distance[40, 41]. Prediction accuracies for short-range base pairs (RNA Contact Distance of 100 or less) for 16S and 23S rRNA comparative structure models in this evaluation were 52% and 50% respectively (Section 2.C.5.1). These results are comparable with previous Gutell Lab studies (Section 2.C.5.1). In contrast, base pairs with RNA Contact Distance greater than 100 were predicted much less accurately for 16S and 23S rRNA, 13% and 24% respectively, (Section 2.C.5.1), and base pairs with RNA Contact Distance greater than 500 were predicted accurately at only 5% (7% if one excludes long-range pseudoknotted base-pairs) and 17% respectively (Section 2.C.5.1).

Beyond absolute prediction accuracy as a function of RNA Contact Distance, one must also consider the distribution of predicted base pairs. Mfold predicts significantly more base pairs then expected for the 16S and 23S comparative structure models in this

evaluation, 17% and 22% respectively (Section 2.C.5.2). Base pairs with a RNA Contact Distance greater than 500 comprise only 2% and 4% of the comparatively predicted base pairs in 16S and 23S rRNA comparative structure models in this evaluation (less if one excludes long-range pseduknotted base pairs in 16S rRNA) (Section 2.C.5.1); however, Mfold predicts significantly more long-range base pairs, 13% and 17% respectively for 16S and 23S rRNA (Section 2.C.5.2). The vast majority of these predicted base pairs are incorrect (Section 2.C.5.2). For the 496 16S rRNA comparative structure models in this evaluation, the base pair prediction accuracy decreases exponentially as the RNA Contact Distance increased (Section 2.C.5.3).

Clearly, given the thermodynamic model and parameters in Mfold 3.1, significantly more long-range pairings are identified as possible when the input is a longer RNA sequences such as 16S and 23S rRNA. Furthermore, one can extrapolate that with shorter RNA sequences the prediction accuracy will be higher due to the fact that: 1) less long range base pairings are possible, and 2) the prediction accuracy for short-range base pairs is the highest in the longer RNAs. This conclusion could be further validated by extending this analysis to the 5S and tRNA comparative structure data sets used in the current evaluation.

3) Inclusion of the suboptimal population demonstrates the potential predictive power of Mfold 3.1 but at the same time provides more evidence that Mfold is unable to accurately and reliably identify the correct RNA secondary structure as the minimum free energy structure. In their evaluation of Mfold 3.1, Mathews et al. state that if one considers a population of 750 suboptimal structure predictions which can vary at most 20% in free energy from the minimum structure prediction then for a given RNA sequence: 1) on average, one of the 750 suboptimal structure predictions contains 86% of expected comparative base pairs and 2) an average of 97.1% of expected comparative

base pairs are identified if one considers the 750 suboptimal structure predictions in-toto. The evaluation in Section 2.C was designed with identifying how well the thermodynamic model reflects RNA secondary structure. Therefore, the set of suboptimal predictions much closer to the optimal structure prediction (within 5%) was analyzed for each of the 496 16S rRNA comparative structure models in this evaluation (Section 2.C.6.1). An average of 71% of the comparatively predicted base pairs are identified when one includes the suboptimal population (Section 2.C.6.1), an improvement in prediction accuracy of 30% on average (Section 2.C.6.1). These results compare favorable with the average accuracy of 77.8% reported by Mathews et al.[44] when the suboptimal population is included. This observation demonstrates the potential predictive power of Mfold 3.1. It can not be overlooked that increase in accuracy is accompanied by a significant decline in recall (Section 2.C.6.1).

If we extend this analysis to consider base pairs with large RNA Contact Distances from the set of 496 16S rRNA comparative structure models in the evaluation in Section 2.C, we find that the number of base pairs with RNA Contact Distance greater than 500 predicted increases significantly; from 193 to 1,460, a 656% increase (Section 2.C.6.2). However, the 63% of the expected comparative base pairs with RNA Contact Distance greater than 500 still never identified (43% if one considers pseudoknots).

SECTION 2.E SUMMARY AND PERSPECTIVE

The current thermodynamic model for an RNA helix was originally proposed in 1972 and first substantiated with experiments in 1974 (Section 2.B). Between 1974 and 1999, a significant amount of work has been dedicated to measuring the energetic parameters for this model using experiments on small RNA oligonucleotides (Section 2.B). This model is based on the concept that the free energy of a helix is a combination of the free energy of initiation (forming the first base pair) and the free energy of

elongation (forming subsequent base pairs). The free energy of initiation includes a destabilizing factor for any loops of unpaired nucleotides which result from the formation of the helix. While the experiments on small RNA oligonucleotides have demonstrated the capacity to provide useful energetic parameters for calculating the free energy of elongation, experimental evidence for calculating the destabilizing contribution of loops has been much harder to establish experimentally, especially for complicated loops such as multistem loops.

The RNA structure prediction evaluation presented in this dissertation (Section 2.C and 2.D) corroborates the claims by Mathews et al. that the improved energy parameters allow Mfold 3.1 to adequately predict RNA secondary structure for shorter RNAs; however, the evaluation also demonstrated a lack of a significant improvement between Mfold 2.3 and Mfold 3.1 in prediction accuracy for longer RNAs such as 16S and 23S rRNA. The lack of significant improvement in secondary structure prediction for longer RNAs contributes clear evidence that the thermodynamic model is still not complete. Furthermore, this evaluation also highlighted a particular RNA secondary structure motif where the existing thermodynamic model does not indicate that the comparatively predicted base pairs are energetically optimal; long-range base pairs and helices. Consider the simple example in Figure 2.9. Two hypothetical 11 base pair helices are represented. In Figure 2.9 (A) the hypothetical helix is capped by a hairpin loop of 200 nucleotides and in Figure 2.9 (B) the hypothetical helix is capped by a hairpin loop of only 10 nucleotides. Computing the free energies of these two helices in an analogous manner to Mfold 3.1 results in free energies of -19.135 kcal/mol for Figure 2.9 (A) and -21.5 kcal/mol for Figure 2.9 (B), with a difference of only 2.365 kcal/mol. To put this into perspective, adding an additional G:C base pair to the helix in Figure 2.9 (B) would

lower its free energy to -22.035 kcal/mol, making it more energetically favorable than the helix in Figure 2.9 (B).

The failure to accurately predict long-range base pairs and helices provides an indication that the thermodynamic model as it is current implemented is not adequate for predicting the secondary structure for larger RNAs such as 16S or 23S ribosomal RNA.. In particular, more energetic parameters are necessary to determine the destabilizing effects of loops, especially multistem loops which are frequently involved in long-range interactions in comparatively predicted RNA secondary structures. Unfortunately, it is very difficult to design RNA oligonucleotides to experimentally determine the energetics of multistem loops. Therefore, one must augment the neareast-neighbor thermodynamic model with constraints from other sources. In particular kinetic experiments and simulations which can provide more insight on folding pathways, and RNA comparative analysis which has already deduced a number of significant constraints (Chapter 3.B.3) are two future sources of constraints and knowledge.

2.F RNA SECONDARY STRUCTURE PREDICTION VIA FREE ENERGY MINIMIZATION SINCE 2003

The evaluation of RNA secondary structure prediction accuracy by Mfold 3.1 presented in this dissertation was conducted in 2002 and published in BMC Bioinformatics in 2004[82]. Since this work was conducted, a considerable amount of research has been conducted on numerous fronts to improve the accuracy of RNA secondary structure prediction by free energy minimization. Many of these results are in line with the conclusions in Section 2.D and 2.F about the necessity to determine more constraints including: 1) The experimental determination of thermodynamic parameters for three and four way multistem loops using optical melting experiments and a new equation for computing multistem loop free energies with an experimental basis[84, 85];

2) Constraints based on chemical modification experiments [86]; 3) The energy parameters revised to accommodate terminal mismatches and additional experiments on bulge loops of a single nucleotides and internal loops[86-89]; 4) Energy parameters to predict the enthalpy change of RNA secondary structure formation; facilitating RNA secondary structure prediction at different temperatures[90]; 5) The dynamic programming algorithm was improved to include coaxial stacking between adjacent helices[86]. 7) Different RNA structure prediction algorithms which can predict pseudoknots, and/or include additional constraints such as common secondary structure, phylogenetic relationships, and other experimental constraints. Mathews and Turner provide a good discussion in[91].

2.G METHODS

2.G.1 Prediction Accuracy Calculations

The accuracy of the structures predicted in this evaluation was scored by quantifying how well the optimal (minimum energy) structure prediction matched the comparative structure model for each sequence in the dataset. Results were only based on sequences folded in their entirety. The accuracy was calculated by dividing the number of comparative base pairs that were predicted exactly with Mfold 3.1 by the total number of canonical base-pairs in the comparative model (excluding any base-pairs with IUPAC symbols other than G,C,A or U). This method for calculating accuracy was the same as the previous Gutell Lab studies[40, 41], with the exception that previous studies excluded comparative base pairs that were pseudoknotted from consideration. In my evaluation, pseudoknotted comparative base pairs were not excluded because it is only due to a limitation of the dynamic programming algorithm and not the thermodynamic model that they are not predicted. From the manual that accompanies Mfold 3.1, "*When pseudoknots*

are included, the loop decomposition of a secondary structure breaks down and the energy rules break down. Although we can assign reasonable free energies to helices in a pseudoknot, and even to possible coaxial stacking between them, it is not possible the estimate the effects of new kinds of loops that are created."

In contrast, in the Mathews et al.[44] study predicted base pairs were considered correct if: 1) they matched a comparatively predicted base-pair exactly or 2) either nucleotide of the Mfold predicted base-pair (X,Y where X and Y are the positions of the nucleotides in the sequence) is within one nucleotide of its comparatively predicted position (X, Y±1 or X±1,Y). While the Mathews et al. study included a measure of the percentage of comparative base-pairs considered pseudoknotted, it was not clear if those base-pairs were specifically excluded from their accuracy calculations. Based on these differences in the accuracy calculations, the Mathews et al. study has the potential to report higher accuracies than the current evaluation.

2.G.2 Per Sequence Averages

Some average values for statistics computed in this evaluation, such as secondary structure prediction accuracy, were calculated on a per sequence basis. A per sequence average variant of a particular statistic was calculated by averaging the value of the statistic for each individual sequence in the dataset. For example, for the 16S rRNA comparative structure model dataset, the overall accuracy was calculated by first determining the accuracy of the Mfold optimal structure prediction for each individual sequence, and then, the 496 accuracy values were averaged to calculate the overall accuracy score of 41%. The Mfold optimal structure prediction accuracy for each of the 1,411 RNA comparative structure models in this evaluation can be found Appendix C.

2.G.3 Logarithmic Binning of Base Pairs by RNA Contact Distance

Figure 2.7 showed that the number of comparative base-pairs observed decreased exponentially as the RNA Contact Distance increased. To determine if this observation was accurate, logarithmic binning was required to group the base-pairs into somewhat equally-sized bins based on RNA Contact Distance. The shortest and longest RNA Contact Distances observed from the 496 16S rRNA comparative structure models were 3 and 1833, respectively. Therefore, the overall range of our logarithmic scale was from log10 (3) to log10 (1833). This range was divided into equal increments to define the RNA Contact Distance bins. After evaluating many increment sets, with the requirement that the sizes of the bins be within one order of magnitude of one another, seven distance bins were established (Figure 2.8). The detailed counts for the 191,994 comparative base pairs from the 496 16S rRNA comparative structure models includes in this evaluation as a function of RNA Contact Distance are available in Appendix D.

Chapter 3: Improving the Efficiency and Throughput of RNA Comparative Analysis via an Expert Systems Approach

3.A INTRODUCTION

In Chapter 2, I established that free energy minimization based on nearest-neighbor energy parameters is still not capable of accurately and reliably predicting the secondary structure of an RNA, although it has improved significantly for shorter RNAs. One important improvement has been the augmentation of the experimentally established thermodynamic model with biases or constraints. These constraints can have multiple sources including: 1) experiments such as chemical modification, 2) RNA structures determined through NMR or X-ray crystallography or 3) RNA comparative analysis. RNA comparative analysis is a knowledge-based methodology for RNA structure prediction as opposed to a rigorous first principles approach such as free energy minimization. The specific application of comparative analysis to RNA involves selecting a diverse set of sequences and comparing them under the assumption that their function is phylogenetically conserved which implies that they share a common structure.

Starting with his involvement as a graduate student in developing the first LSU and SSU Ribosomal RNA secondary structure models[50, 51], Dr. Robin Gutell has been actively developing systematic methods to extend and improve the comparative analysis of the LSU and SSU Ribosomal RNA and Group I and II Introns. In the process, Dr Gutell realized the importance of efficient data organization and management in the comparative analysis process. Dr. Gutell founded the CRW Project[62] dedicated to the comparative analysis of all Ribosomal RNA, Group I and II introns and Transfer RNA sequences identified in public sequence repositories. Dr. Gutell and co-workers created an organized and systematic data management infrastructure which included a database

management system and disseminated the results of their comparative analysis to the scientific community though a web-based presentation. To ensure high accuracy and quality, Dr. Gutell has been an advocate of an expert systems approach to RNA comparative analysis. In an expert systems methodology, the biologist is intimately involved in the analysis of the data, and software tools are used to assist rather than replace the biologist.

Since 2000, the number of RNA sequences available to the CRW Project has expanded rapidly. As a result, the CRW Project has been under significant pressure to improve its data curation processes to accommodate this large increase while maintaining its high standards for the accuracy and quality of its analysis. *As a graduate student in Dr. Gutell's lab, my primary research emphasis has been on analyzing the CRW Project data management infrastructure and engineering software tools to enhance and improve the expert systems approach to RNA comparative analysis. Individual algorithm development is important, and many computer scientists, bioinformatics researchers and mathematicians focus exclusively on this aspect; however, I would contend that building vertically integrated analysis infrastructures which provide for significant quality control and draw from disparate, non-connected data sources is just as biologically significant and computationally challenging.* The CRW Project amongst others has already demonstrated the importance of this approach. In this chapter I discuss one significant result of my efforts, the Comparative Analysis Toolkit (CAT). CAT as it exists today provides a solid software foundation for developing more enhanced tools and techniques to promote the expert systems approach to RNA comparative analysis. In the remainder of this chapter I provide a summary of the history of RNA comparative analysis as applied to Ribosomal RNA; followed by a discussion of the expert systems methodology of the CRW Project and the design and implementation of CAT.

3.B HISTORICAL PERSPECTIVE AND SCIENTIFIC BACKGROUND

As discussed in Chapter 1, the concept that an RNA can form secondary structure in-vivo that is a collection of helices which consist of intramolecular interactions was established in 1959[1]. The tRNA secondary structure was posited by comparing the first three tRNA sequences determined by enzymatic digestion[45, 46]. The remainder of this section will focus primarily on the comparative analysis of the Ribosomal RNAs (5S, 16S and 23S), as RNA comparative analysis was defined and enhanced through research with these particular RNAs.

3.B.1 RNA Comparative Analysis to Establish the first Secondary Structure Models for 5S, 16S and 23S rRNA

In 1975, Fox and Woese were the first to utilize the concept of phylogenetic conservation of functionally significant features in predicting RNA secondary structure for 5S rRNA[49]. They posited that functionally equivalent molecules will have similar secondary structure. From the set of all possible helices predicted with free energy minimization for 5S rRNA[27, 29, 31], Fox and Woese assembled a secondary structure model for 5S rRNA which contained only phylogenetically conserved helices.

Subsequently in 1980 when the first 16S and 23S rRNA sequences were obtained[92-94], comparative analysis was again used to predict a minimal secondary structure for 16S[50] and 23S rRNA[51]. Both 16S and 23S rRNA were assumed to form a phylogenetically similar secondary structure. The initial structure models were posited from an alignment of two 16S and 23S rRNA sequences and a set of phylogenetically diverse 16S and 23S rRNA oligoribonucleotides respectively. These initial rRNA sequences were similar enough to be aligned by identity, but had enough variation to identify common structure. Columns in the alignment involved in a structural relationship were determined through visual inspection for coordinated, compensating changes

between canonical base-pairs as illustrated in Figure 3.1; the concept was termed *positional covariation*. These minimal structure models required helices to have a minimum of four base pairs. Chemical modification and enzymatic digestion experiments were performed to validate the predicted structure models. By 1983, a larger, phylogenetically diverse set 16S rRNA sequences was available. These additional sequences were manually aligned and subsequently used to refine the 16S rRNA secondary structure model[54, 95]. The requirements for predicting a helix were modified to require that its existence be established by comparative evidence including: 1) it must be present in at least two different 16S rRNA sequences and 2) *positional covariation* could be established for at least two base pairs in the helix.

3.B.2 Developing Tools and Methods to extend RNA Comparative Analysis

Initially, the rRNA sequence alignments were edited manually in the default UNIX text editor *vi*. Starting in 1986, Tom Macke in Dr. Carl Woese's lab developed the multiple sequence alignment editor, AE1. Between 1987 and 1992 Macke in conjunction with Dr. Gutell improved AE1 and AE2 was developed. AE2 was implemented in C based on the UNIX *curses* library. AE2 has its own ASCII-based format for RNA sequence alignments and does not provide support for input and output in any other common bioinformatics formats. RNA secondary structure diagrams were created with the first interactive RNA secondary structure drawing program, *Stred*. *Stred* was developed by Bryn Wiser in Dr. Harry Noller's lab in conjunction with Dr. Gutell. It was subsequently re-developed in C for the X-windows environment using the Motif widget library and was named XRNA[96]. Similar to AE2, XRNA has its own ASCII-based format for RNA secondary structures. Since 2004, XRNA has been available as a Java-based package.

Beyond tools for manipulating and visualizing RNA sequence alignments and secondary structure diagrams, it was necessary to develop systematic methods for RNA comparative analysis. In 1985, a method for covaration analysis termed the *number-pattern* method was developed and applied to a large alignment (> 30) of 16S rRNA sequences[52]. The basic concept of the *number-pattern* method was to convert sequence patterns for a column of the alignment into numerical patterns which could be easily sorted and compared with UNIX tools like *sort*. Columns with similar number patterns were candidates to be base paired. In 1986, tracking phylogenetic events spanning the three major domains of the Tree of Life were used to support a tertiary interaction (570:866 base pair in 16S rRNA, E.coli numbering) initially predicted with covariation analysis[97]. In 1992, a rigorous statistical method, *MIXY*, was developed to identify columns which have similar patterns of variation[53]. *MIXY* has no built in bias for canonical Watson-crick base-pairs, and considers the nucleotide frequencies from all columns of the alignment equally.

3.B.3 The Application of RNA Comparative Analysis to Larger Sets of 16S and 23S rRNA Sequences

As a result of the development of both software tools and systematic algorithms for RNA comparative analysis, even more divergent 16S and 23S rRNA sequences were aligned manually in AE2, guided by known patterns of conservation and structural constraints. *MIXY* analysis was used to identify columns of the alignment with common patterns of variation, and the results were used to refine and improved the alignment in an iterative manner. As a result of the expansion of the RNA sequence alignments, many new RNA structural features were discovered including: dominant G:U base pairs[98], G:A mismatches[99], other non-canonical base-pairs[55, 100], lone pairs[55], base triples[101], tetraloops[78, 79, 102] and pseudoknots[55, 100].

Following the expansion of the RNA sequence alignments, the first databases of 16S and 23S rRNA secondary structure models predicted with comparative analysis and drawn with XRNA were made available[57-59]. Secondary structure models present in these databases spanned the entire Tree of Life and included both Mitochondrial and Chloroplast in addition to Nuclear-encoded sequences. Concurrently, other research teams also published secondary structure models for 16S and 23S rRNA using comparative analysis techniques[103-106] and databases of 23S rRNA secondary structure models[107].

The existence of large accurate and phylogenetically diverse alignments of 16S and 23S rRNA sequences; large databases of comparatively predicted 16S and 23S rRNA secondary structure models; the growing corpus of RNA structural features discovered through RNA comparative analysis; and high resolution crystal structures of the yeast phenylalanine tRNA[47, 48], hammerhead ribozyme[108], and the P4-P6 domain of the *Tetrahymena thermophilia* Group I Intron[13] provided a foundation for deciphering more complex higher order structural constraints associated with RNA structure including: tetraloop receptors and tertiary interactions involving tetraloops[77, 109, 110], unpaired adenosines in the covariation-based structure model[52, 111], A:A and A:G oppositions/base pairs at the ends of helices[54, 112, 113], E loops/S turns[111, 114, 115], E-like loops[111], adenosine platforms[111, 116], A-minor motif[117, 118], Kink-turn[119, 120], U-turns[121], lone-pair triloops[122, 123], and the UAA/GAN internal loop motif[124]. The CRW Project has developed an interactive Java applet, RNAMap (Gandhi et al., unpublished data), which allows a user to visualize these different motifs overlaid on a two-dimensional comparatively predicted structure model. RNAMap is currently available at the CRW Web site[62].

3.B.4 The Accuracy of the rRNA Comparatively Predicted Structure Models as Validation of RNA Comparative Analysis

In 2000, the high resolution crystal structures for the 30S and 50S ribosomal subunits were determined[60, 61], an extremely important achievement in modern molecular biology. These experimentally determined, high-resolution structures provided a comprehensive vehicle for assessing the accuracy of the 16S and 23S comparatively prediction secondary structure models. Approximately 97-98% of the base pairings present in the covariation-based secondary structure models for 16S and 23S rRNA were found in the high resolution crystal structures[25]. The majority of predicted base-pairings that were not in the crystal structure were A:U and G:U pairings with minimal covaration, but were included in the model because they occurred at the ends of helices. All base pairs with extensive covariation and comparative support were identified in the crystal structures[25]. Furthermore, all tertiary and tertiary-like base-pairings that were predicted with covariation analysis were also present in the crystal structures. These results confirm more than just the accuracy of the comparatively predicted secondary structure models; they validate: 1) the fundamental premise underlying RNA comparative analysis, the phylogenetic conservation of functional elements; 2) the process of aligning RNA sequences based on structure; 3) using covaration analysis to identity positions within the sequence which exhibit similar patterns of variation; 4) refining those alignments through interpretation of the results of covariation analysis.

3.B.5 CRW Web Site and Project: Curation and Dissemination of RNA Comparative Analysis Artifacts

The usefulness of the large 16S and 23S rRNA comparative structure databases disseminated by the Gutell Lab[56-59] prompted the development of a more sophisticated data management infrastructure to collect **and** comparatively analyze all 16S, 23S and 5S rRNA's, the Group I and II introns and tRNA into a single, self-

consistent data collection under strict standards for accuracy and quality control. The data collection is disseminated to the scientific community via a large web site; the CRW Web Site and Project[62]. The CRW Web Site was first available to the public in January 2000 and provided a significant number of useful RNA comparative analysis artifacts including: 1) A web-based interface to querying a relational database that contains metadata about the RNA sequences collected and analyzed by the CRW Project, the CRW Project RNA Metadata Database. Relevant metadata included the phylogenetic designation, NCBI accession identifier and the computed percent complete and length after alignment; 2) Direct access to over 400 comparatively predicted RNA secondary structure diagrams for 16S rRNA, 23S rRNA, 5S rRNA and Group I and II introns; 3) The manually curated and refined RNA sequence alignments used to predicted the comparatively predicted structures; 4) Detailed nucleotide frequency tables that demonstrated the conservation and variation of sequence elements across the Tree of Life for 5S, 16S, 23S rRNA, tRNA and Group I and II Introns; 5) Base pair frequencies for the 16S and 23S covariation-based secondary structure models which demonstrate the conservation and variation of secondary structure elements across the Tree of Life; 6) a new covaration analysis algorithm named *Covary* and confidence rankings for predicted base pairs; 7) secondary structure "conservation diagrams" for 16S and 23S rRNA which display the conservation and variation of both sequence and structure elements for sequences on different parts of the Tree of Life (Figure 2.1 and 2.6). Complementary comparative sequence and structure artifacts are available from other web sites. Several of these projects RDP[125], Greengenes[126] and the European rRNA Database[127] are more narrowly focused, catering primarily to a phylogenetic audience. Others have a primarily structural slant such as Rfam[128]. Only the CRW Web Site and Project addresses both the phylogenetic and biophysical aspects of RNA comparative analysis.

In the following sections I discuss the development of the vertically integrated data management and curation infrastructure at the CRW Project. The development of this infrastructure is based on the success of the expert systems approach to RNA comparative analysis as demonstrated from the history comparative analysis of Ribosomal RNA. The guiding lesson from this history is that the intuition of the biologist is necessary for accurate and meaningful results. RNA comparative analysis software should be developed to compliment rather than replace the biologist. The key to successful RNA comparative analysis is the generation of the proper alignment of the RNA sequences. The expert systems approach to RNA comparative analysis used by the CRW Project is based on a theoretical model for the architecture of an RNA sequence alignment. The model is used to classify RNA sequences into one of four categories and define suitable methodologies for analyzing RNA sequences that fall within those categories. In Section 3.B.5.1, I discuss the theoretical model for the biological architecture of the RNA sequence alignment. In Section 3.B.5.2 I introduce a methodology for analyzing RNA sequences from an expert systems perspective based on this methodology. Finally, in Section 3.B.5.3 I introduce the initial implementation of the expert systems approach to RNA comparative analysis used by the CRW Project, the *Curation Pipeline*.

3.B.5.1 The Architecture of an RNA Sequence Alignment from a Theoretical Perspective

A fundamental assumption of RNA comparative analysis is that a given RNA type has a phylogenetically conserved function conferred through its structure (Section 3.B.1). Consequently, one can analyze divergent RNA sequences of the same type by assuming that they share a common or core structure. Of course, the problem is complicated by the fact that for a given RNA type, not all organisms form the same **exact**

structure, but only share some common core structure. The success of RNA comparative analysis has been that it implicitly searches for and utilizes the complex relationships between sequence, structure and phylogeny that are embedded within the RNA sequence alignment. In this section, I examine the theoretical architecture of an RNA sequence alignment and the relationships between sequence, structure and phylogeny which governs the alignment

First, I define a theoretical model for an RNA sequence alignment with the assumption that the total sequence space for a given RNA type is not infinite. In Figure 3.2 I present an abstract sequence space plot representing a hypothetical RNA sequence alignment for a given RNA type which contains a sampling of sequences spanning the entire Tree of Life. In Figure 3.2, a few "islands" of the total RNA sequence space have been sampled rather extensively (labeled A, B and C, Figure 3.2). RNA sequences within any of these "islands" are closely related, exhibit high levels of sequence and structure conservation and can be aligned rather easily with one another based on sequence. However, when one compares sequences between different "islands" (e.g., a sequence in "island" A compared with a sequence in "island" C), more sequence variation is observed. One can identify common patterns of variation between these "islands". These patterns of variation are used to subsequently align these "islands" with one another. The common patterns of variation identified between the islands represent common structure. Figure 3.3 is a schematic of an RNA sequence alignment in a form that is generally recognizable, as a matrix. Different "islands" of sequence space identified in Figure 3.2 are indicated by differential coloring. The columns which represent core common structure (surrounded by black vertical boxes in Figure 3.3) exhibit a common pattern of variation between the "islands."

To provide circumstantial evidence for this theoretical description of the architecture of an RNA sequence alignment I analyzed the 16S rRNA alignment of 6326 complete sequences spanning the entire Tree of Life, available at the CRW Web Site[62]. The accuracy of this alignment has been established through the comparison of predicted secondary structure model for 16S rRNA to the high resolution X-ray structure (Section 3.B.4). For each pair of sequences in this alignment, I compute the "Phylogenetic Distance", the "Sequence Identity" and the "Structural Identity". "Phlyogenetic Distance" is defined as the number of links on the phylogenetic tree between two sequences (Figure 3.4). The "Sequence Identity" is the percentage of overlapping nucleotides that are identical (Figure 3.5). The "Structural Identity" is defined as the percentage of overlapping nucleotides to the total number of columns in the alignment (Figure 3.5). In Figure 3.6 I plot the average "Sequence Identity" between pairs of sequences against the "Phylogenetic Distance", and in Figure 3.7 I plot the average "Structural Identity" between pairs of sequences against the "Phylogenetic Distance". Sequences with a "Phylogenetic Distance" of 0 have and average "Sequence Identity" of 95% while sequences with a "Phlyogenetic Distance" of 8 have and average "Sequence Identity" of c.a. 83% (Figure 3.6). In contrast, sequences with a "Phylogenetic Distance" of 0 have an average "Structural Identity" in excess of 98%, while sequences with a "Phylogenetic Distance" of 8 still have an average "Structural Identity" in excess of 96%. The results are consistent with the idea that sequences within "islands" exhibit high "Sequence Identity" and "Structural Identity" and are closely related. Sequences between "islands" are more distantly related, exhibit lower "Sequence Identity" but high "Structural Identity." In fact, the "Structural Identity" remains in excess of 90% for sequences with a "Sequence Identity" of 65% or higher (Figure 3.8). These relationships are implicitly

represented on the 16S rRNA secondary structure conservation diagrams available from the CRW Web Site (Figure 2.1)[62].

3.B.5.2 Developing a Strategy for an Expert Systems Approach to RNA Comparative Analysis

Based on this theoretical architecture of an RNA sequence alignment introduced in Section 3.B.5.1, one can build a "divide and conquer" methodology for analyzing RNA sequences with comparative analysis by segmenting newly identified sequences into one of four categories. Category 1 sequences are newly identified RNA sequences that fall within an "island" that have been identified and therefore should be expected to conform to the known structural constraints (Figure 3.9). To align sequences in this category, all that is required is the identification of the appropriate "island" that the sequence belongs to within the existing RNA sequence alignment and that once the sequence is aligned within that region, no known patterns of variation are violated.

Category 2 sequences are newly identified RNA sequences that do not have sufficient sequence identity with sequences in any of the "islands" identified in the existing RNA sequence alignment nor do they have enough identity with one another to be aligned into their own island. In order to align these sequences accurately and reliably, more sequences must be collected such that these individual sequences can be grouped into "islands". Once an "island" has been identified, the sequences within the "island" can be confidently aligned with one another. Subsequently, common patterns of variation can be identified to reconcile these new "islands" with existing "islands." Furthermore, these new "islands" may also have entirely new patterns of variation unique to that island. Automated methods have been developed to solve the alignment of Category 2 sequences without waiting for additional samples; however, their accuracy and quality are not up to the standards of the CRW Project. I address these methods in Section 3.D.

Category 3 sequences are RNA sequences that exhibit a significant amount of identity with sequences within an existing "island" of the alignment; however, they may contain regions of "hypervariability." The sketch in Figure 3.10 is an example of a region of "hypervariability" within an existing "island." Biologically, regions of "hypervariability" are large insertions within a couple of sequences that have been grouped into an "island." The alignment of these hypervariable regions can not be solved until more sequences are identified that are part of the "island" and contain this insertion.

Category 4 sequences fit into none of the previous three categories, even after one significantly increases their sampling of RNA sequence space. Possible biological explanations for these sequences include: 1) errors in sequencing have occurred and the given sequence is not an RNA sequence of the type annotated, 2) this sequence is on the edge of acceptability as a biologically viable candidate of an RNA sequence of the type annotated and closely related sequences occur in organisms that are not viable. This argument is strengthened if the particular RNA of interested is vital to survival of the cell such as Ribosomal RNA.

3.B.5.3: The Curation Pipeline: The CRW Web Site and Project Data Management and Analysis Infrastructure

Developing the initial data management infrastructure for the CRW Project was a significant undertaking which involved: 1) efficiently organizing data such that it can be quickly retrieved and analyzed, 2) developing methodologies and software tools for analyzing large numbers of RNA sequences by comparative analysis, and 3) developing methodologies for verifying the accuracy of the comparative analysis and any artifacts generated. Figure 3.11 is a graphical depiction of the data management infrastructure and the analysis pipeline developed initially by the CRW Project, guided by the theoretical picture of the architecture of an RNA sequence alignment introduced in Section 3.B.5.1,

and the "divide and conquer" methodology introduced in Section 3.B.5.2. Unaligned RNA sequences which fit into Category 1 and Category 3 were the initial focus. The implementation of this pipeline was via complex maze of UNIX and Perl scripts, standalone C++ programs and a significant amount of manual, repetitive curation work, both in the data analysis and the data management. Throughout the rest of this chapter, I refer to this pipeline as the *Curation Pipeline*. A summary of the major stages in the *Curation Pipeline* follows (Figure 3.11).

Stage 1: RNA sequences of interest are identified in Genbank[129] through an annotation-based search. At the beginning, this was done manually by Dr. Gutell, but in late 2002/early 2003 was replaced with an automated process capable of: 1) automatically identifying any Ribosomal RNA sequence from the Genbank daily update and 2) checking the CRW Project RNA Metadata Database to determine if a given sequence had already been collected. As the identification is based on annotation, it is not exhaustive and consequently relevant RNA sequences are missed because they are incorrectly annotated or not annotated at all. Furthermore, few of these annotation errors are sufficiently systematic enough to make automated resolution viable. The CRW Project is in the process of developing better methods to identify RNA sequences of interest in Genbank. One area of promise is using programs such as RNAMOT[130, 131], RNAMotif[132], and ERPIN[133] to identify sequences that match defined or inferred structural descriptors.

Stage 2: RNA sequences identified in Stage 1 are manually sorted by a CRW Project Biologist (Figure 3.11). Each RNA sequence is excised from its Genbank entry according to the annotation. Many times, extra nucleotides are excised on the 5' and 3' end due to high probability of mis-annotation. These extra nucleotides are addressed in Stage 4. The excised RNA sequence is inserted into the proper *holding alignment*. A

holding alignment is an ASCII-based AE2 formatted text file that contains only unaligned RNA sequences. For example, all 16S rRNA sequences identified are placed in the 16S rRNA holding alignment. In contrast, all aligned RNA sequences are kept in the *main alignment* for a given RNA type. The *main alignment* is also an ASCII-based AE2 formatted text file. Concurrently, a new entry is made in the CRW Project RNA Metadata database for the excised RNA sequence which includes the Genbank identifier of the particular sequence (Figure 3.11). Finally, the excised RNA sequence is assigned a unique name to facilitate subsequent manipulation.

Stage 3: A given RNA sequence is aligned in two distinct steps. The first step is a rough sequence-based alignment guided by a sequence that has already been aligned; this step is the focus of Stage 3. The CRW Project originally developed a C++ program named "*autoalign*" to rough align a given RNA sequence using a sequence that is already aligned as a "template". The CRW Project biologist is responsible for selecting the appropriate "template" sequence, and "template" selection is done manually, in some cases with the assistance of ad-hoc FASTA[134-136] queries of the *main alignment*. The "*autoalign*" program is designed to only align an RNA sequence in regions of high identity with the selected "template". Aligning the sequence in more variable regions is left to the biologist to do manually, primarily by using existing secondary structure constraints. Variable regions with no identified structural constraints remain unaligned until enough closely related RNA sequences are identified such that the variable region can be solved iteratively using covariation analysis and manual juxtaposition. Once the common patterns of conservation and variation in a variable region are determined, the newly aligned sequences can be used as a templates to align other RNA sequences in the former variable region. This concept is depicted pictorially in Figure 3.10.

Beyond aligning a given RNA sequence in regions of high identity, "*autoalign*" **attempts** to insert within the alignment of the sequence gap characters and other annotation symbols which were embedded in the *main alignment*. The purpose of these additional annotation symbols is to make it easier to visualize secondary structure helices and other structural elements within the AE2 editor. For example, a 16S rRNA sequence of 1500 nucleotides could easily have another 3000 to 4000 additional annotation characters inserted within it once it was introduced into the alignment. Inserting and maintaining these annotation characters manually would entail a significant amount of manual work. It must be noted that many times, the current "*autoalign*" program does a poor job of inserting these symbols correctly and the biologist is left resolve this manually in Stage 4. The aligned and annotated version of the RNA sequence is output in AE2 format.

Stage 4: The second step in the sequence alignment process is a manual refinement and cleanup; this step is the focus of Stage 4. The "*autoalign*" result is imported into the *main alignment* using the AE2. The CRW Project biologist manually completes the alignment of the RNA sequence in AE2. This involves: 1) correcting any parts of the alignment that "*autoalign*" solved incorrectly and 2) identifying and fixing any errors in the insertion of annotation characters. For variable regions of the sequence which "*autoalign*" could not solve, the biologist can attempt to manually solve the alignment to the extent that secondary structure constraints are available. If the region is too variable, the biologist can not align it until more closely related relatives become available (Figure 3.10). To facilitate this manual refinement process, the CRW Project biologist manually sorts the sequences within the *main alignment* by their known phylogenetic relationships. The CRW Project biologist may employ iterative covariation analysis and manual juxtaposition as suggested in Step 3.

The CRW Project biologist manually verifies that the newly aligned RNA sequence satisfies known pattern of sequence conservation and variation. This check is necessary to ensure that all sequences within the *main alignment* remain properly aligned with one another. Once this check is complete and the alignment of the given RNA sequence is determined to be acceptable, the proper ends of the RNA sequence are determined manually within the context of the alignment, and the computed percent complete and length are manually entered into the CRW Project Metadata Database for the given RNA sequence. The re-determination of the sequence ends is necessary because many times, the ends of a given sequence in the Genbank entry are annotated incorrectly. Therefore, when an RNA sequence is identified, "extra" nucleotides beyond the annotated 5' and 3' ends are excised in Step 2. The given RNA sequence is manually renamed in a more descriptive manner indicating its phylogenetically assigned scientific name. This name is entered in the CRW Metadata database to maintain the link between the alignment and the database created in Step 2.

Stage 5: If appropriate, a secondary structure diagram is generated for the sequence using a program written by Nan Lin[137] using an existing secondary structure diagram as a template. The biologist manually refines the generated diagram in XRNA and enters it into the CRW Project Metadata Database as tentative. Secondary structure diagrams are not generated for every sequence, just representative nodes of the phylogenetic tree. A secondary structure diagram is reviewed for accuracy and quality before being made available to the public.

The CRW Project *Curation Pipeline* described above is a continuum. At any point in time, different sets of RNA sequences are at different stages in the pipeline. For Stages 3 and 4 each RNA sequence identified by the CRW Project is currently analyzed individually. This *Curation Pipeline* has been developed with the primary goal of

maintaining standards for accuracy and quality while attempting to automate some of the more repetitive parts of the pipeline. However, to handle more variable regions of the sequence alignment process the intuition of the CRW Project biologist is still a necessary element for successful and meaningful results. Therefore, the *Curation Pipeline* is not completely automated and provides ample opportunity for the CRW Project biologist to be involved in the analysis.

3.B.6 CRW Web Site and Project: Growth in the Size of the RNA Sequence Collections

Advances in sequencing technology which resulted from the human genome project have lead to exponential growth in the number of sequences available in Genbank[129]. Between 1998 and 2006, the total number of sequences has increased from 2 million to ~91 million; a 4500% increase[138]. Between 1998 and 2002, Genbank grew from 2 million sequences to 22 million sequences[76]. Beyond just sequences, many complete genome sequences for organisms spanning the entire Tree of Life are now available. As of March 2007, 471 complete Microbial (36 Archaea and 435 Bacterial), 26 complete Eukaryotic and 1115 complete Animal Mitochondrial Genomes were available[138].

As the Ribosomal RNA (rRNA) is the most popular gene for phylogenetic analysis to determine relationships between different organisms on the Tree of Life, the number of rRNA sequences identified in Genbank is expected to grow at a similar rate. Table 3.1 provides a summary of: 1) the number of RNA sequences analyzed by the CRW Project by July 2003; where an analyzed sequence is defined as having progressed through all 5 steps of the *Curation Pipeline* (Figure 3.11), 2) the number of RNA sequences identified in Genbank by the CRW Project via annotation-based searches through July 2003, and 3) the number of rRNA sequences identified in Genbank by the

CRW Project through March 2007. In July 2003 79,250 RNA sequences had been identified, but only 26,600 had been analyzed (Table 3.1) and by March 2007 a total of 973,400 RNA sequences had been identified (Table 3.1), more than a 10 fold increase. The stated goal of the CRW Project is to analyze all RNA sequences of interest identified in the public sequence repositories. One can extrapolate from these figures that the manually intensive RNA comparative analysis *Curation Pipeline* as outlined in Section 3.B.5.3 will not be able to scale to accommodate such a large increase in the number of RNA sequences available to analyze. To accommodate this large increase in the number of RNA sequences available for analysis, it is necessary to increase the throughput of the *Curation Pipeline* at the CRW Project.

3.C THE COMPARATIVE ANALYSIS TOOLKIT (CAT): A SOFTWARE TOOLKIT TO STREAMLINE THE RNA COMPARATIVE ANALYSIS *CURATION PIPELINE*

In Section 3.B, I briefly discussed the history of RNA comparative analysis from the perspective of the analysis of Ribosomal RNA. This history culminated with the launch of the CRW Web Site and Project in 2000 to analyze and align all Ribosomal RNA and Group I and II intron sequences identified in Genbank[129] while maintaining strict standards for accuracy and quality. The CRW project utilizes an expert systems approach to RNA sequence alignment and by extension RNA comparative analysis (Section 3.B.5). To implement their methodology in a systematic and rigorous manner, the CRW Project has developed a data analysis and management workflow, the *Curation Pipeline* (Section 3.B.5.3). However, the CRW Project *Curation Pipeline* as discussed in Section 3.B.5 is far from optimal. Table 3.2 depicts the number of unaligned RNA sequences identified by the CRW Project at each major stage in the *Curation Pipeline* in July 2003. As expected, the largest bottlenecks are the semi-manual sequence alignment process (Stage 3, Section 3.B.5.3) and the manual clean-up, refinement and verification

process (Stage 4, Section 3.B.5.3) which also happen to be the most biologically challenging and interesting. In order to meet the ambitious goals of the CRW Project, it is necessary to develop better RNA comparative analysis tools to streamline the *Curation Pipeline* by automating manual, time-consuming repetitive tasks and more efficiently utilizing the existing corpus of comparative data available including the CRW Project RNA Metadata Database, manually curated RNA sequence alignments, and comparatively predicted RNA secondary structure models.

In late 2003/early 2004, as a graduate student in the Gutell Lab the development of the software tools to facilitate the expert systems approach to RNA comparative analysis became the primary emphasis of my research. I developed the Comparative Analysis Toolkit (CAT) software package which: 1) provided a modern software solution to address bottlenecks in the *Curation Pipeline* and 2) created a software foundation for the development of future tools for expert systems based RNA comparative analysis. CAT is a platform-independent, modular command-line based application that supports a scriptable, unattended batch mode as well as an interactive mode and integrates directly with the CRW RNA Metadata Database. CAT contains both biologically novel software methods as well as more straightforward utilities for streamlining individual data manipulation tasks within the *Curation Pipeline*. The first release of CAT within the Gutell Lab was in late 2004. Since then, I have been actively maintaining and incrementally enhancing CAT. As of March 2007, the current version of CAT is 0.2.22.5, and contains in excess of 100,000 lines of C++ and Java code. Some of the most biologically novel features such as the "*evaluator*" module (Section 3.C.3) have been developed only within the last 18 months or so.

While CAT integrates in one software package a large number of different utilities and commands for streamlining different data manipulation and data analysis

tasks within the *Curation Pipeline*, the discussion of CAT in this dissertation will be limited to the most biologically significant elements. These elements include: 1) redeveloping "*autoalign*" (Section 3.C.2.1) to be more efficient about inserting and maintaining annotation characters within the alignment of a given RNA sequence, 2) more fully automating structure-based RNA sequence alignment by providing semi-automated, higher-throughput methods for "template" sequence (Section 3.C.2.3 and 3.C.3.4) and 3) developing automated methods for evaluating the quality of a structure-based RNA sequence alignment using known sequence and structure constraints, and indicating to the CRW Project Biologist where focus manual refinement effort (Section 3.C.3). Strategically, addressing these elements would provide the biggest productivity boost to the CRW Project in the near-term by addressing unaligned RNA sequences that fall into Category 1 or Category 3 (Section 3.B.5.2) by utilizing the existing corpus of comparative data already available. Throughout the rest of this section I discuss the specific elements of CAT. Areas for future development of CAT are deferred to Chapter 4. The entire user manual for CAT through version 0.2.21 is available at the CRW Web Site[62].

3.C.1 CAT: Architecture

CAT is a platform-independent software foundation for the expert systems approach to RNA comparative analysis. Figure 3.12 is a simple block architecture diagram of CAT. The CAT program is a designed to receive input from disparate data sources including: 1) the CRW RNA Metadata Database, 2) Genbank, 3) the Biologist, 4) RNA secondary structure models in BPSEQ and XRNA formats[62], and 5) AE2-formatted RNA sequence alignments (Figure 3.12). Outputs of the CAT program include: 1) AE2- or FASTA-formatted RNA sequence alignments, 2) RNA Secondary Structure Models in XRNA format and 3) SQL commands to directly update the CRW RNA

Metadata Database as a result of comparative analysis carried out within CAT (Figure 3.12).

Since the number of RNA sequences the CRW Project intends to analyze was already known to be in excess of 10^5 sequences and growing rapidly, and that fact that the greater scientific community lacks tools for RNA comparative analysis from an expert systems perspective, CAT was implemented as a Java/C++ layered application in order to: 1) ensure good performance when manipulating large RNA sequence alignments which can contain 10^5 sequences and 2) to provide a cross-platform implementation such that CAT could eventually be distributed to the greater scientific community. (**Note:** Distributing CAT to the scientific community is not a primary focus of this dissertation.) As indicated in Figure 3.12, the user interface and the computational engine are implemented in Java while the in-memory alignment data structures are implemented in C++. CAT uses the Java Database Connectivity (JDBC) library within Java to interact with the CRW RNA Metadata Database and the NCBI E-Utils XML-based Web Services interfaces[138] to interact directly with Genbank[129]. CAT can automatically "*fork*" child processes using the Runtime.*exec* function, which is a native part of the Java Programming Language. CAT uses this functionality to automatically invoke FASTA[134-136] as part of automated "template" sequence selection for "*autoalign*" (Section 3.C.2.3 and 3.C.2.4) In the rest of this section, I briefly discuss some of the implementation features.

3.C.1.1 A Novel, High-performance C++ In-Memory Alignment Data Structure

A primary requirement for CAT was the ability to manipulate RNA sequence alignments that will eventually grow to be in excess of 10^5 sequences. This capability sets CAT apart from other software applications developed for biological sequence manipulation. This requirement presented an interesting design challenge: how does one

engineer a modular, scalable and computationally intensive cross-platform application with extremely large in-memory data structures? While the Java Virtual Machine (JVM) confers many benefits (Section 3.C.1.2), its major drawback is that the user does not have explicit control over the allocation and de-allocation of memory within a given application process; this is always handled by the JVM. Since an RNA sequence alignment of 10^5 sequences with $5x10^3$ columns has a potential memory footprint of 3.7GB, robust performance requires that the CAT be directly responsible for allocating and managing this memory space. Therefore, the native in-memory RNA sequence alignment data structure is implemented as a native C++ object. This object is allocated on the memory heap managed by the CAT application (Figure 3.13). A corresponding Java RNA sequence alignment object created within the JVM as a proxy (Figure 3.13). The Java RNA sequence alignment proxy object holds a pointer to the location of the corresponding native C++ object on the CAT application heap (Figure 3.13).

Figure 3.14 is a Unified Modeling Language (UML) object diagram that depicts the different C++ and Java objects which together constitute the in-memory alignment data structure. When CAT is required to load a particular RNA sequence alignment into memory, a new *Alignment* object (Figure 3.14) is instantiated on the CAT application heap and the corresponding alignment is loaded into that object. The corresponding Java *AlignmentProxy* object is created and is retained within the user-interface layer (Figure 3.14). There is a one to one correspondence between an *Alignment* object and an *AlignmentProxy* object as illustrated by the multiplicities in Figure 3.14. I should point out that CAT is capable of holding multiple RNA sequence alignments in memory simultaneously as illustrated by the multiplicities on the object association between the *CATApplication* object and the *AlignmentProxy* object (Figure 3.14); one *CATApplication* object can be associated with zero or more *AlignmentProxy* objects. All

function calls within the Java user-interface layer of CAT which utilize the in-memory RNA sequence alignment are made on the *AlignmentProxy* object. Using the pointer to the native C++ object held by the *AlignmentProxy* object and JNI, the calls are forwarded to the appropriate C++ RNA sequence alignment object on the heap, executed and the resulting values returned to the user (Figure 3.15). This implementation retains platform-independent characteristics because the native C++ alignment data structure can easily be implemented in ANSI C++ and therefore can be compiled on UNIX/Linux, Windows and Mac OSX.

The software design of the C++ in-memory RNA sequence alignment data structure itself was motivated by performance concerns of manipulating a large two-dimensional in-memory array of 10^8 elements (10^5 sequences by 10^3 columns). In particular a novel indirection mechanism was designed to facilitate inserting columns within the large two-dimensional memory array without having to re-arrange the entire data structure each time. Figure 3.16 is a schematic of the design and provides an example of how column insertion and deletion are streamlined by the indirection. The actual ordering of the columns in the alignment is maintained with a simple one-dimensional array of *AlignmentColumn* objects (Figure 3.16). Each *AlignmentColumn* column object can point to any column of the allocated two-dimensional memory array (Figure 3.16). The only requirement is that no two *AlignmentColumn* objects point to the same column of the two-dimensional array. When a new column is inserted, only the one-dimensional array of *AlignmentColumn* objects is modified to reflect the insertion. The new *AlignmentColumn* object representing the inserted column uses the first available column in the two-dimensional array (Figure 3.16), which can theoretically be anywhere in the array. Of course, deleting a column is simply a matter of removing the particular *AlignmentColumn* object from the one-dimensional array. The column in the two-

dimensional array is marked as available and can be utilized when the next column is added (Figure 3.16).

3.C.1.2 Java-based Command-Line Driven User Interface Layer

The CAT user interface layer is command-line driven and implemented in Java. Figure 3.17 is a screen capture of the CAT interface with a single RNA sequence alignment loaded in-memory. Many advantages are conferred by using Java for the interface implementation in particular: 1) platform-independence, 2) increased stability over an entirely native C/C++ implementation, and 3) rapid program development due in part to a large library of utility classes which are provided as part of the Java programming language.. Complex platform specific issues such as threading and file system management are abstracted by the Java Virtual Machine (JVM). Within CAT, RNA comparative analysis computational algorithms are implemented in a robust, multi-threaded without any concern for platform-specific thread management issues.

The programmatic implementation of the user interface is highly object-oriented and modular. Programmers can easily add new commands to the core CAT application. Figure 3.18 is a UML object diagram depicting the main Java class hierarchy and inheritance relationships for the different objects which comprise the user interface. One of the key goals in the design of CAT was to promote reuse. I was recognized early on that many of 40 commands to be implemented as part of CAT used components of other commands internally. Therefore, the CAT user interface layer was designed to support *command chaining*. Figure 3.19 is a UML sequence diagram which provides an example of *command chaining*. "Command Chaining" promotes re-use and reduces the maintenance overhead by concentrating functionality in single, re-usable blocks (i.e., commands). CAT provides other user-interface features that are useful for accomplishing specific tasks within the *Curation Pipeline* such as: 1) command aliasing and 2) un-

attended scriptable batch execution mode. Since the software architecture of CAT from a user-interface perspective is beyond the scope of this dissertation the reader is directed to CAT Users Guide available at the CRW Web Site [62]to learn more about the design and implementation of CAT.

3.C.2 CAT: Semi-Automated RNA Sequence Alignment with an Enhanced "a*utoalign*" Module

The original "*autoalign*" program was developed by the CRW Project as a simple method of bootstrapping the manual RNA sequence alignment process in Stage 3 of the *Curation Pipeline* (Section 3.B.5.3) given a manually identified "template" sequence which is: 1) closely related to the unaligned RNA sequence, 2) already aligned and 3) representative of the patterns of sequence conservation within the existing RNA sequence alignment. Category 1 and/or Category 3 unaligned RNA sequences (Section 3.B.5.2) which can be grouped into islands of closely related sequences which exhibit high sequence and structural identity are the most amenable to this strategy. Within CAT, two automated RNA sequence alignment strategies have been developed that combine "template" sequence selection and rough alignment with completely re-designed and re-developed "*autoalign*" algorithm. These strategies systematically utilize: 1) phylogenetic relationships that are stored in the CRW Project RNA Metadata Database and 2) the rough sequence and structural identity, as computed with FASTA[134-136], between an unaligned RNA sequence and a candidate "template" sequence.

The intent in creating automated RNA sequence alignment strategies within CAT was to streamlining the expert systems approach to RNA comparative analysis used by the CRW Project by using our knowledge of the architecture of the RNA sequence alignment to our advantage (Section 3.B.5.1). They are not intended to replace the CRW Project Biologist in the *Curation Pipeline*, and they are not immediately useful for

automatically aligning Category 2 sequences which potentially represent new "islands" of sequence space (Section 3.B.5.2). The CRW Project biologist must identify and align sequences within the "island" before the automated sequence alignment strategies in CAT are useful. In this section I discuss the complete re-design and re-development of the "*autoalign*" methodology in CAT followed by the two automated sequence alignment strategies and their implementation in CAT.

3.C.2.1 Enhanced "autoalign" Algorithm

The first version of the "*autoalign*" algorithm developed by the CRW Project simply looked for consecutive matches of 10 or more between a template sequence and an unaligned RNA sequence, any shorter regions of overlap between the "template" and the unaligned RNA sequence were ignored. In contrast, the "*autoalign*" algorithm implementation in CAT uses a heuristic, recursive pairwise approach to align RNA sequences given the identification of a suitable "template" sequence. The pairwise alignment methodology can be classified in the family of Needleman and Wunsch[139], Sankoff[140] and Smith and Waterman[141] with one important caveat: "*autoalign*" does not compute a rigorously maximal sequence alignment between the unaligned RNA sequence and the specified "template" using gap penalties. The reason for this is that the mandate of "*autoalign*" within the expert systems approach to RNA comparative analysis is only to **bootstrap** the alignment process by solving regions of high sequence and structure identity between the unaligned RNA sequence and the "template" sequence. In other words, "*autoalign*" is designed to align an RNA sequence into an existing "island" of high sequence and structural identity (Section 3.B.5.1), The hypothesis is that if an unaligned RNA sequence can be grouped into an existing "island" and a "template" sequence from that "island" identified to guide the alignment of the unaligned RNA

sequence, then the amount of manual alignment work left for the CRW Project biologist will be minimal.

Throughout the remainder of this discussion, the unaligned RNA sequence is referred to as the "query" sequence. The heuristic pairwise alignment procedure in "*autoalign*" begins by computing a dot plot between the "query" and "template" sequence to identify the longest line of similarity (Figure 3.20) (A). Because "*autoalign*" assumes that the "query" and "template" sequences exhibit high sequence and structural identity, dot plot computation begins from the center diagonal (y-intercept = 0) and proceeds out towards the edges (Figure 3.20) (A). By default, "*autoalign*" computes only 10% of the dot plot, based on its assumptions of high sequence and structural identity between "query" and "template". Specific examples where this heuristic can lead to incorrect results are: 1) when the "query" sequence is a partial sequence which cannot be expected to have high structural identity with the "template" and 2) if the query has a large insertion relative to the "template" (Category 3, Section 3.B.5.2). To handle these cases, the "*autoalign*" algorithm adjusts the dot plot computation extent based on the length difference between the "query" and "template" sequences (Figure 3.20) (B). The basic formula for dot plot computation extent (DPE) is $DPE = m\Delta L_{tq} + DPE_{\min}$ where $\Delta L_{tq} = \left|L_t - L_q\right| \div L_q$ if $L_q > L_t$ or $\Delta L_{tq} = \left|L_t - L_q\right| \div L_t$ if $L_t > L_q$. ΔL_{tq} is the length difference between the "query" sequence and the "template" sequence, L_t is the length of the "template" sequence and L_q is the length of the "query" sequence. By default m is 0.9 with a DPE_{min} of 0.10 (10%). The use of a partial dot plot to identify the maximal line of similarity is the primary reason that "*autoalign*" is not rigorous.

The longest line of similarity is selected from the partial dot plot (Figure 3.21) (A), and extended in either direction as long as: 1) the number of mismatches between two co-linear similarity lines is below a certain threshold and 2) the length of the

adjacent, co-linear similarity line is above a specified threshold (Figure 3.21) (B). This approach is able to quickly identify the gross alignment between the sequences (by identifying the large regions of similarity) while still tolerating small regions of variation. Even though the "query" and "template" exhibit high sequence and structural identity, some variation is expected, especially in regions that covary (Section 3.B.5.1). When the selected similarity line can no longer be extended, new dot plots are computed at either end of the selected similarity line and the process is repeated (Figure 3.22) (A). The algorithm continues recursively until no more lines of similarity can be identified (Figure 3.22) (B). The difference between the first step and subsequence recursive steps in "*autoalign*" is that no dot plots computed in any recursive steps are partial. This approach only computes a single alignment solution between the "query" and the "template" sequence. Furthermore, it is not guaranteed to be the maximal alignment; however, the mandate of "*autoalign*" is to bootstrap the alignment process, not solve it rigorously.

Once the alignment between the "query" and "template" is computed, a column mapping is generated between the "query" and the "template" based on the computed alignment and the alignment of the "template" sequence within the existing RNA sequence alignment (Figure 3.23). For example, if nucleotide 8 in the "query" sequence is aligned to nucleotide 9 in the "template" sequence, then nucleotide 8 in the "query" sequence should be assigned to the same column that nucleotide 12 in the "template" sequence currently occupies (Figure 3.23). Once this column mapping is generated "*autoalign*" has a separate algorithm for translating the newly aligned query sequence into the existing RNA sequence alignment based on this column mapping while simultaneously inserting all necessary annotation symbols.

This algorithm can be considered a sliding window approach which starts from the 5' end of the existing RNA sequence alignment. Following the defined column

mapping between the "query" and the "template", nucleotides in the "query" sequence are placed into the appropriate columns within the alignment as defined by the mapping (Figure 3.23). Any annotation symbols from the alignment of the "template" are inserted in corresponding columns in the alignment of the "query" when those columns are not occupied by a nucleotide of the query sequence (Figure 3.23). When the number of nucleotides intervening between two mapped nucleotides in the "query" sequence is less than the number of columns, the algorithm will insert as many query nucleotides as possible, and then consider itself to be *out of register* (Figure 3.23). Once the algorithm is *out of register*, the computed alignment mapping is ignored and nucleotides are inserted sequentially within columns of the existing RNA sequence alignment until the algorithm can get back into register (Figure 3.23). Any nucleotides mapped while the algorithm is out of register are lower case for quick visual inspection. The algorithm is not allowed to rectify an *out of register* situation by adding columns into the alignment because it is assumed that the computed alignment may not be optimal, and the CRW Project biologist should inspect it. The final, translated result for the "query" is temporarily held in-memory, and the CRW Project biologist can select other "template" sequences to align "query" against. Furthermore, the CRW Project biologist can use the "*evaluator*" (Section 3.C.2) to score the different alignments for the "query" if they chose multiple "templates" while the results remain in-memory. Once the user has finished aligning a sequence with "*autoalign*" and evaluating the results, they can export the best result to an AE2 formatted file and import it directly into the main alignment (Section 3.B.5.3).

3.C.2.2 The Performance of the Enhanced "autoalign" Algorithm

To assess the productivity of this newer version of "*autoalign*" we must consider each of its primary functions independently. First, I should characterize the accuracy of the alignment of a given "query" sequence to a "template" sequence using the recursive

dot-plot approach methodology. The accuracy is characterized as a function of the identity between the "template" sequence and the "query" sequence and is reported in Table 3.3. The alignment accuracy is directly related to the sequence identity between the "template" and the "query." When the sequence identity between the "template" and the "query" is estimated to be at least 91%, the alignment accuracy is at least 92% (Table 3.3). Second, I should characterize the efficiency of the translation of the "query" sequence into the existing alignment given the mapping computed by "*autoalign*". Figure 3.24 is a plot of the total number of errors for a translated "query" sequence as a function of the total number of columns in the alignment. When the "template" and the "query" exhibit a sequence identity of 85% or higher, the total number of columns in error in the final translated alignment result is less that 3% (Figure 3.24). From these results we observe that when a given unaligned RNA sequence exhibits high sequence and structural identity with its template, "*autoalign*" efficiently maps the unaligned RNA sequence into the existing RNA sequence alignment.

3.C.2.3 "Full Alignment" Semi-Automated Sequence Alignment Module

The "*Full Alignment*" module in CAT can be considered a top-down or rigorous, semi-automated alignment strategy. "*Full Alignment*" can be summarized as: given an unaligned RNA sequence and an alignment of existing RNA sequences, rigorously search the existing RNA sequence alignment for all suitable sequence(s) which can be used as templates to align the given unaligned RNA sequence with *autoalign*; invoking the *autoalign* command to align the unaligned RNA sequence using each candidate "template" sequence identified; invoking the "evaluate" command (Section 3.D.2) to check the accuracy of each possible alignment result for the unaligned RNA sequence and selecting the best alignment. The strategy is summarized with the UML sequence diagram in Figure 3.25. Currently, the "*Full Alignment*" implementation in CAT does not

automatically rank and select the best alignment for an unaligned RNA sequence; the CRW Project biologist still makes this determination, using the results of the evaluator to guide their decision.

By default, the "*Full Alignment*" module searches the entire existing RNA sequence alignment using FASTA[134-136]. Only aligned sequences that have a minimum identity (default 95%), a minimum overlap (default 90%), and a maximum length difference (default 25%) with the unaligned RNA sequence as per FASTA[134-136] are considered as candidate "template" sequences. The minimum identity parameter is used to enforce sequence identity while the minimum overlap parameter is used to enforce structural identity between the unaligned RNA sequence and any candidate "template" sequence identified. The minimum length difference parameter prevents a partial sequence from selection as an *autoalign* "template" for a complete sequence. Furthermore, the extent of the search for suitable "template" sequences from the existing RNA sequence alignment can be limited by optionally specifying a maximum "Phylogenetic Distance" (Section 3.B.5.1) between the unaligned RNA sequence and any candidate "template" sequence. If "Phylogenetic Distance" constraints are specified, the implementation of "*Full Alignment*" in CAT first directly queries the CRW RNA Metadata Database for the phylogenetic placement of all sequences in the existing RNA sequence alignment. Next, "*Full Alignment*" computes the "Phylogenetic Distance" between each sequence in the existing RNA sequence alignment and the unaligned RNA sequence. The subset of sequences from the existing RNA sequence alignment that are within the specified "Phlyogenetic Distance" from the unaligned RNA sequence are selected. Subsequent steps in the methodology are only applied to the phylogenetically constrained subset.

The biggest downside to the "*Full Alignment*" methodology is that each unaligned RNA sequence must still be processed individually. However, by automating the manually intensive step of selecting the best suitable "template" sequences, the CRW Project biologist is only left to evaluate the results of aligning the given unaligned RNA sequence using *autoalign* and each selected "template" sequence. The evaluator module (Section 3.C.3) is designed to assist the CRW Project biologist in this process. The "*Full Alignment*" automated sequence alignment strategy significantly reduces the amount of manual work required by the CRW Project biologist in Stage 3 of the *Curation Pipeline* (Section 3.B.5.3).

3.C.2.4 "Find Queries" Semi-Automated Sequence Alignment Module

Contrary to the rigorous top-down approach in "*Full Alignment*", "*Find Queries*" uses the opposite strategy and can be considered an optimistic or "bottom-up" semi-automated alignment strategy. "*Find Queries*" can be summarized as: given a single, aligned RNA sequence from an existing RNA sequence alignment as a "template" and a set of unaligned RNA sequences, search the set of unaligned RNA sequences to identify all unaligned RNA sequences which can be aligned using *autoalign* and the specified "template" sequence; invoke the *autoalign* command to align each unaligned RNA sequence identified against the specified "template" sequence; invoke the "evaluate" command to check the accuracy of alignment for each unaligned RNA sequence identified. With "*Find Queries*", the CRW Project biologist is only presented with a single alignment result for each unaligned sequence. The strategy is summarized in the UML sequence diagram in Figure 3.26.

In similar manner to "*Full Alignment*" (Section 3.C.2.3), "*Find Queries*" uses a FASTA[134-136] search to identify unaligned RNA sequences with minimum identity (default 95%), minimum overlap (95%) and a maximum length difference (10%) to the

specified “template” sequence. The biological significance of each criterion is the same for “*Full Alignment*” and “*Find Queries*” Similar to the “*Full Alignment*” approach, the number of candidate unaligned sequences to search can be restricted by specifying a maximum “Phlyogenetic Distance” between the specified “template” sequence and any unaligned sequence.

The “*Find Queries*” approach is more efficient than the “*Full Alignment*” approach because it identifies all unaligned sequences for which the given aligned sequence is a suitable “template”, instead of repetitively using “*Full Alignment*” on each unaligned sequence. However, on the downside, the CRW Project biologist is presented with only one possible alignment for each unaligned RNA sequence. This approach is more risky from an accuracy perspective if the unaligned RNA sequences identified can not be accurately grouped into the same “island” with the selected “template” and therefore the criterion are more stringent.

3.C.3 CAT: Automated Alignment Evaluation Module

Within Stage 4 of the *Curation Pipeline*, the accuracy and quality of the alignment of an RNA sequence is currently evaluated manually (Section 3.B.5.3). One area where CAT can significantly improve the expert systems approach to RNA sequence alignment and RNA comparative analysis is by providing an automated alignment “*evaluator*” module which uses both sequence-based and structure-based techniques to characterize alignment quality. This “*evaluator*” module should have the capability to indicate to the CRW Project Biologist in which areas of the alignment they should focus their manual clean-up efforts, and should compute the actual percent complete and length of given RNA sequence within the context of the total RNA sequence alignment.

In the discussion of the architecture of an RNA sequence alignment we demonstrated that different “islands” within sequence space can be identified where all

sequences within the “island” exhibit significant sequence and structural identity once they are aligned with one another (Section 3.B.5.1). Once consequence of this phenomenon is that the number of degrees of freedom within the alignment for that “island” are significantly decreased and alignment errors can be easily identified using conserved sequence and structure constraints that apply to that “island”. For example, Figure 3.27 (A) is a subset of the well-characterized alignment of five sequences from the same “island” of the bacterial segment of the 16S rRNA alignment[62] with the 5’ and 3’ halves of a secondary structure helix consisting of five base pairs identified. Figure 3.27 (B) is a “diff view” of this “island” where all positions with a nucleotide equivalent to the nucleotide in the first row are displayed with a ‘.’. From this “diff view” we observe that the sequences all exhibit a significant amount of sequence identity with one another. If we insert an extra gap character ‘-‘ into the third sequence in the block at column 9, 20 exceptions become visible in the “diff view” (Figure 3.27). Furthermore, if we consider the base pair defined between columns 9 and 25 in the helix marked in Figure 3.27 (A), we see that it is a G-C for all five rows in the block. When the extra gap character is inserted in Figure 3.27 (C), the third sequence in the block no longer retains the base pairing relationship for columns 9 and 25.

This example qualitatively illustrates both a sequence-based and a structure-based approach to evaluating the quality of the alignment of a sequence. A computationally efficient way to implement sequence-based evaluation is via a consensus sequence, which is a method of summarizing the patterns of nucleotide conservation and variation within a block of sequences in an RNA sequence alignment. The failure of a given sequence to conform to the consensus in highly conserved regions of an RNA sequence alignment can be a significant indicator of mis-alignment. Figure 3.28 illustrates the 90% consensus for a block of 24 sequences from the bacterial segment of the 16S rRNA alignment[62]. The

consensus is computed by analyzing the nucleotide frequencies in each column individually. The *level* of the consensus (e.g., 90% in Figure 3.28) represents a threshold frequency for analyzing each column. For the consensus to represent a given column as having a nucleotide, a percentage of rows greater than or equal to the *level* must have a nucleotide in that column. For the consensus to represent a given column as having a specific nucleotide (e.g., a 'G'), a percentage of the rows greater than or equal to the *level* must have the specific nucleotide in that column. Figure 3.28 is depicts the consensus computation for two columns in a hypothetical RNA sequence alignment.

In an analogous manner, a computationally efficient method to implement structure-based evaluation is via computing base-pair frequencies which summarize patterns of secondary structure conservation and variation within a block of sequences in an RNA sequence alignment that share a common structure. The absence of conserved structural features can be significant indicator of mis-alignment. Computing base pair frequencies is simply a matter of computing nucleotide frequencies across two columns in an RNA sequence alignment simultaneously. Figure 3.29 is a simple example of computing the base pair frequencies for three columns from an RNA sequence alignment used in Figure 3.28. Beyond just computing the overall base-pair frequencies, computing frequencies for different aggregations such as the percentage of Watson-Crick base pairs is also useful, this is due to the fact that the conservation of base pair type (e.g., G-C, A-U, etc) is not a direct indicator of the conservation of a particular base pair. In fact, identifying covariations is the method by which common secondary structure is deduced (Section 3.B.1).

The discussion in this section is intended to introduce the novel "*evaluator*" module developed within CAT. The "*evaluator*" utilizes the sequence-based and structure-based techniques discussed above to determine the quality of the alignment of a

given RNA sequence within an existing RNA sequence alignment. Initial evaluation results can be further analyzed to identify regions of significant mis-alignment ("hotspots") and the percent complete and length of a given RNA sequence can be computed within the context of the RNA sequence alignment. The results of the evaluation are delivered to the user as a summary report in tabular format, suitable to be imported into spreadsheet programs such as Microsoft Excel for further analysis. The "evaluator" module is also capable of generating AE2 formatted alignment annotations to graphically indicate evaluation results to the CRW Project biologist.

3.C.3.1 Sequence-based Evaluation

The goal is to quantify from a sequence perspective how well the alignment of a given RNA sequence agrees with alignment of other sequences with which it shares an "island" (Section 3.B.5.1). The task can be divided into three steps. The first step is to identify a set of closely related RNA sequences are that have already been determined to be accurately aligned within the existing RNA sequence alignment. Criterion used in the selection include: 1) metadata from the CRW RNA Metadata Database such as phylogenetic classification and percent complete (Section 3.B.5); 2) specific sequence and/or structural identity to a given reference sequence. The second step is to compute a consensus sequence using the sequences selected in the first step. An example consensus sequence computation was provided in the introduction to Section 3.C.3. In the third step, the alignment of RNA sequence in question is compared against the computed consensus on a column by column basis. The five most important categories defined to summarize the comparison in any individual column between the consensus and the alignment of the RNA sequence in question are: 1) (**cM**) exact IUPAC nucleotide to IUPAC nucleotide match, 2) (**cFUZ**) a non-specific nucleotide in the consensus to IUPAC nucleotide, 3) (**cMIS**) mismatch IUPAC nucleotide to IUPAC nucleotide, 4) (**cNQA**) specific or non-

specific IUPAC nucleotide in the consensus to a gap, 5) (**cAQN**) gap or annotation in the consensus to IUPAC nucleotide. An example is provided in Figure 3.30. The column index of each mismatch or fuzzy match (**cMIS** or **cFUZ**) is output to a text file enabling the CRW Project biologist to quickly navigate to these positions in the alignment with AE2. By default the "*evaluator*" module applies a consensus level of 90%. The default consensus level was selected based on the data presented in Section 3.B.5.1 which indicated that sequences within the same "island" exhibit significant sequence identity with one another.

A sequence that is well-aligned within a given "island" should have a high number of matches, either exact matches (**cM**) or fuzzy matches (**cFUZ**) combined with a small number of mismatches (**cMIS**). Furthermore, since the consensus is computed over a set of well-aligned sequences that exhibit significant sequence and structural identity, one can argue that the number of fuzzy matches should small. Consider a hypothetical RNA sequence alignment in which the sequences are all at least 90% identical with one another. The 90% consensus will well-defined with few opportunities for fuzzy matches exist as the sequences exhibit high identity. By comparison, a hypothetical RNA sequence alignment in which the sequences can have as little as 70% identity with one another have a 90% consensus is significantly more non-specific as the sequences exhibit significantly more sequence variation. Beyond just considering matches and mismatches to the consensus, the sum of the gaps between the consensus and the particular sequence evaluated (**cNQA** and **cAQN)** is another indicator of mis-alignment. The sequences selected for the consensus are expected to share an "island" with the sequence under evaluation; therefore, they should exhibit a significant amount of structural as well as sequence identity. The presence of a significant number of gaps is a metric which suggests significant structural variation.

3.C.3.2 Structure-based Evaluation

Our goal is to quantify from a structural perspective how well the alignment of a given RNA sequence agrees with alignment of other sequences with which it shares an "island" (Section 3.B.5.1). The first step is to introduce a set of structural constraints into CAT for a given reference sequence. In the current implementation of the "*evaluator*" module, these constraints are limited to secondary structure base-pairings which are represented as binary relationships between different columns within the CAT in-memory alignment data structure (Figure 3.16). The second step is to compute base pair frequencies by projecting the secondary structure pairings introduced into CAT across the same set of closely related and well-aligned RNA sequences selected for sequence-based evaluation (Section 3.C.3.1). An example base pair frequency computation was provided in the introduction to Section 3.C.3. In the last step, the secondary structure constraints introduced in the first step are projected across the alignment of the RNA sequence under evaluation to determine: 1) the presence or absence of expected base pairs and 2) the extent of agreement between base pairs that are present and the patterns of structural conservation and variation characterized by the computed base pair frequencies. The three most important categories defined to summarize the comparison between any individual base pair present and the expected patterns of conservation and variation for that base pair are: 1) (**sBPD**) the base pair formed matches the dominant or most frequently occurring base pair; 2) (**sBPM**) the base pair formed matches a base pair which occurs above a specified threshold frequency; 3) (**sBPWC**) the base pair formed is a Watson-Crick base pair and Watson-Crick base pairs occur above a specified threshold. The expected values for (**sBPM**) and (**sBPWC**) are computed by comparing the overlap between the alignment of the reference sequence for which the base pair relationships were specified and the alignment of the RNA sequence under evaluation. Figure 3.31 is

an example of computing these three metrics for a given RNA sequence and a specified set of secondary structure base pairings projected on an existing RNA sequence alignment. As part of this evaluation, it is possible to identify mismatches and fuzzy matches (**cMIS** and **cFUZ**) from the sequence-based evaluation (Section 3.C.3.1) which should be considered to be aligned correctly when known secondary structure constraints are considered.

Since sequences within an "island" are expected to exhibit significant structural identity as well as sequence identity (Section 3.B.5.1), a sequence that is well-aligned within a given "island" should have values for (**sBPM**) and (**sBPWC**) that are very close to the expected values. This follows from the basic premise of the "*evaluator*" module, that well-characterized "islands" within an existing RNA sequence alignment exhibit very few degrees of freedom. Because the sequence identity is expected to be high, the structural consensus will be well-defined. The number of base pairs which are either exactly the same or simply a Watson-crick exchange (e.g., G:C <=> A:U) will be high when the secondary structures between any two RNA sequences within the "island" are compared with one another. It is important to note that where sequence-based evaluation is only applicable within a specific "island" of the RNA sequence alignment; as a result of the principles of phylogenetic conservation of structure and *positional covariation* (Section 3.B.1), structure-based evaluation can be used check the accuracy of the alignment of entire "islands" with one another. Two "islands" in the RNA sequence alignment that exhibit significant sequence variation with one another still have the potential to exhibit a common pattern of variation in columns of the alignment which map to known secondary structure constraints (Section 3.B.5.1). Figure 3.32 is a simple schematic that depicts how structure-based evaluation can be used to assist in the alignment multiple "islands" with one another. In this example, "islands" 1 and 3 are well

aligned with one another, but "island" 2 is out of register. In this simple example, the absence of expected base pairs in "island" 2 would alert the CRW Project Biologist that it is not properly aligned with the rest of the "islands."

3.C.3.3 Identifying Regions of Significant Alignment Errors

One area of specific focus with this first generation "*evaluator*" module is to identify significant regions of mis-alignment or "hotspots". Efficient detection of "hotspots" has the potential to significantly increase the sequence curation rate at the CRW Project because in most cases they can be manually corrected quickly or the "*autoalign*" result can be scrapped and a new "template" selected. Many times, "hotspots" arise from simple errors where one or two nucleotides are out of register in the alignment, resulting in a large number of downstream errors. Figure 3.27 is a simple example of a one nucleotide error in highly conserved block of a Bacterial 16S rRNA sequence alignment.

The "*evaluator*" module includes a separate analysis component, the "*MismatchAccumulationAnalyzer*." This component scans the sequence-based evaluation result for a given RNA sequence and cumulatively tracks matches, mismatches and fuzzy matches (Section 3.D.3.1). Each column where the accumulated number of mismatches is above the specified threshold is reported to the CRW Project biologist. The "*MismatchAccumulationAnalyzer*" starts at the first column in the RNA sequence alignment with an accumulator value of zero. It then proceeds to examine each column in the alignment individually. For a given column, if a mismatch is detected, the value of the accumulator is incremented by a specific amount. Once a mismatch is detected, each subsequent match or fuzzy match detected until the next mismatch results in a decrementing of the accumulator. The accumulator can not be decremented below zero. By default, the accumulator is incremented by a value of 1 when a mismatch is

encountered, decremented by a value of 0.75 when a match is encountered after a mismatch and 0.50 when a fuzzy match is encountered after a mismatch. Since the goal is to detect significant regions of misalignment, consecutive mismatches are weighted to carry a significant penalty which requires a significant number of matches to rectify. An example of the "*MismatchAccumulationAnalyzer*" is provided in Figure 3.33.

When the "*MismatchAccumulationAnalyzer*" is requested as part of the evaluation of the alignment of a given RNA sequence, an additional column is reported in the results summary labeled (**Analyzer**). The value reported is the number of columns for which the accumulator matched or exceeded the specified threshold value given the sequence-based evaluation of the alignment of the RNA sequence in question. Additionally, each column index where the accumulator matches or exceeds the specified threshold is reported to the Biologist; facilitating a quick assessment of the regions of the alignment in question.

3.C.3.4 Computing the Percent Complete for a RNA Sequence

In the Stage 2 of the CRW Project *Curation Pipeline*, when a given RNA sequence identified through an annotation-based search of Genbank is identified, extra nucleotides are excised beyond the annotated 5' and 3' ends due to the potential for mis-annotation (Section 3.B.5.3). Examples of mis-annotation include: 1) failure to annotate the 5' and/or 3' ends correctly or 2) the failure to annotate one or more introns within the RNA sequence. One that sequence has been aligned within an existing RNA sequence alignment, the 5' and 3' ends can be determined more confidently. It is only at this point that the CRW Project determines the percent complete and length for a given RNA sequence and enters those values into the CRW Project RNA Metadata Database. In the description of the *Curation Pipeline* in Section 3.B.5.3, the determination of the percent complete and length for a given RNA sequence is determined manually in Stage 4.

The "*evaluator*" module can compute the percent complete and length of a given RNA sequence utilizing the consensus sequence computed in the sequence-based evaluation step (Section 3.C.3.1). The fact that the same consensus used in the sequence-based evaluation step is also used for the percent complete computation is important because that consensus is expected to adequately represent the patterns of sequence conservation which should be exhibited by the RNA sequence of interest. Figure 3.34 provides a simple schematic on how the percent complete is computed. For any sequence, the total number of nucleotides which overlap the consensus is considered to be the *Effective Length*. For example, Sequence 2 in Figure 3.34 has an Effective Length of 21 nucleotides where Sequence 4 has an Effective Length of 13 nucleotides. The actual lengths of Sequence 2 and 4 are 32 and 19 respectively (Figure 3.34). To compute the percent complete, the *Effective Length* of a sequence is divided by the number of nucleotides in the consensus. In the example in Figure 3.34, Sequence 2 has a percent complete of 100 (21/21) while Sequence 4 has a percent complete of 62 (13/21).

3.D Other Algorithms and Strategies for Fully-Automated RNA Sequence Alignment

The most difficult challenge in RNA comparative analysis is creating an accurate RNA sequence alignment. This problem can be enormously complex. The number of possible juxtapositions for a pair of RNA sequences of length 1000 is $10^{767.4}$[142]. To complicate matters, two Ribosomal RNA sequences of the same type but from different organisms on the Tree of Life can have as little as 30% identity with one another. The solution to the RNA sequence alignment problem developed in CAT is a multifaceted approach which involves both semi-automated sequence alignment and automated alignment evaluation. However, an area of significant research in bioinformatics and computational biology has focused on developing fully-automated algorithms for

aligning homologous DNA, RNA and protein sequences. In this section, I explore the automated RNA sequence alignment algorithms starting with sequence-based pairwise alignment algorithms and then moving to two distinct classes of multiple sequence alignment algorithms progressive and probabilistic.

The first automated sequence alignment algorithms developed were pairwise algorithms. Their goal was to find the maximal set of non-overlapping identical subsequences between two sequences. The first algorithms were based on the methods of Needleman and Wunsch[139], Sankoff[140] and Smith and Waterman[141] which rigorously compute the maximal set of identical subsequences using dynamic programming algorithms. The equivalence of nucleotides is determined using matrices of acceptable substitutions such as PAM [143] or Blosum 62[144]. Substitution matrices are developed from an analysis of an alignment of properly aligned sequences that are all within a given evolutionary distance of one another. Gaps occur in the alignment when identical subsequences between two sequences can not be identified. Figure 3.35 represents an alignment of identical subsequences connected by gaps. Different penalties are applied for both opening gaps and extending gaps. While these algorithms are guaranteed to find the optimal alignment between two sequences based on the matrix of acceptable substitutions and the set of gap penalties specified, they are not computationally efficient when the requirement is to search a large sequence database for the sequences identical to a given search sequence. Heuristic approaches were developed to improve the computational efficiency such as FASTA[134-136] and later BLAST[145, 146]. FASTA and BLAST derive a significant performance advantage over more rigorous pairwise alignment algorithms by not computing and analyzing all possible subalignments. In the case of FASTA the user specifies the “ktup” parameter or word

size. FASTA will only consider subalignments between the sequences that are as long as or longer than the specific word size.

Beyond pairwise alignment algorithms another class of algorithms was developed for multiple sequence alignment called progressive algorithms. Clustal[147-150] and T-Coffee[151, 152] are the most popular members of this class of algorithms. These algorithms build a multiple sequence alignment by taking advantage of the fact that sequences are evolutionarily related and that identity between closely related sequences is higher than identity between more distantly related sequences (Section 3.C.1). Clustal first computes the pairwise alignment between all pairs of sequences, creating a distance matrix. From that distance matrix a guide tree is created and sequences are progressively aligned according to the guide tree. Pairwise alignment in Clustal is either via heuristic methods or dynamic programming using gap penalties and substitution matrices.

Clustal is remains one of the most popular programs for multiple sequence alignment of proteins, despite that fact that it performs much worse than other programs[153]. Given that Clustal is expected to perform better for protein sequence alignment compared with nucleic acid sequence alignment, one should not hold high hopes that Clustal will work well for RNA sequence alignment. In fact, the accuracy of Clustal declines rapidly as the identity between the sequences decreases. Figure 3.36 (Gutell and Eargle unpublished results) depicts the accuracy of Clustal as a function of sequence identity for all sets of pairwise alignments as compared to the manual structure-based alignment for a set of 800 small subunit animal mitochondrion rRNA sequences. Pure sequence-based methods fail to align more divergent RNA sequences accurately because they have no concept of conserved secondary structure. For an analogy with our biological RNA sequence alignment architecture in Section 3.B.5.1, these methods work well within an “island”, but can not be used to aligned sequences between “islands.”

Because additional constraints such as common secondary structure are necessary to align more divergent RNA sequences, a new class of probabilistic alignment algorithms were developed[154, 155] based on formal language theory. A formal language is simply a set of strings, and a grammar is a precise description of a formal language. Grammars can be broken up into two categories, generative and analytical. Generative grammars are simply a set of rules known as *productions* for generating strings in a language. A given grammar consists of: 1) a finite set of nonterminal symbols, 2) a finite set of terminal symbols disjoint from the set of nonterminal symbols and 3) a finite set of production rules which map one string of symbols to another and the first string contains at least one non-terminal symbol. Consider the following grammar with two production rules: S => aS | S => b. This simple grammar can makes strings of repeating a's terminated by with a single b such as: *aab*, *aaaab*, etc. When the left hand side of all production rules in a grammar contains only a single non-terminal symbol, it is considered to be 'context-free.' The simple example I gave above is a context-free grammar. Context-free grammars have significant performance advantages over regular grammars. Context-free grammars can be combined with probability weightings which incorporate both sequence constraints and secondary structure constraints to create a probabilistic model (known as a stochastic context-free grammar) which can be applied to RNA sequence alignment. The probabilities are either provided *a priori* or can be inferred from observing known RNA sequence alignments. Once a model is determined, aligning any RNA sequence simply requires aligning that sequence to the model and computing the probability that the sequence fits the model. To align a given RNA sequence to the model, different parse trees are constructed according to the grammar (many times these grammars are ambiguous which is why different parse trees can be constructed) and a dynamic programming algorithm is used to select the best parse tree.

This parse tree represents the alignment of the given RNA sequence to the grammar. The probability that the given RNA sequence fits the grammar is the sum of the probabilities of all possible parse trees.

Unlike sequence-based methods, I don't have any data to explicitly disqualify automated sequence base methods based on stochastic, context-free grammars as viable techniques for fully automating RNA sequence alignment within the CRW Project *Curation Pipeline* other than: 1) these alignment algorithms are unable to properly handle pseudoknots, and 2) for best results the probabilistic model must be specified *a priori* which is dependent upon the manual identification of the different patterns of variation that link different "islands" within the RNA sequence alignment. While recent implementations have overcome the pseudoknot limitation[156] we have decided that scaling up the expert system strategy originally developed by the CRW Project (Section 3.B.6) is the best first step. Down the road, once we have identified a large number of "islands" within the RNA sequence alignment, we intend to revisit probabilistic methods based on stochastic grammars.

3.E SUMMARY AND PERSPECTIVES

The expert systems approach to RNA comparative analysis developed by the CRW Project and the development of the Comparative Analysis Toolkit (CAT) to enhance, refine and streamline this approach in light of the large increase in diversity and volume of RNA sequences identified has been the focus of this chapter. I began this chapter with a brief discussion of the history of RNA comparative analysis as applied to studying rRNA. In this time period, rRNA sequences were determined and collected which spanned the entire Tree of Life. Methods were developed to analyze these sequences from a comparative perspective and deduce common secondary structure models. Experimental evidence was collected to verify the models, and in 2000, the first

complete high-resolution crystal structures of the LSU and SSU rRNAs were published, validating the comparative techniques used to predict the rRNA secondary structure models in the intervening years. In this time period, many other RNAs beside rRNA were studied from a comparative perspective. The large corpus of RNA comparative data collected has provided significant insight into the fundamentals of RNA structure. Many higher order sequence-structure motifs have been first identified from observed biases in the comparative data and later verified experimentally and/or through detailed analysis of the crystallographic data.

To maintain the RNA comparative data in a form suitable to facilitate scientific analysis and discovery, the CRW Project was established. With the CRW Project, Dr Robin Gutell applied the unique insights and knowledge gained from his direct involvement in the pioneering work of the previous 20 years in the development of an expert systems approach to RNA comparative analysis, the *Curation Pipeline*. The *Curation Pipeline* is designed to facilitate the systematic comparative analysis of RNA sequences given the biological architecture of the RNA sequence alignment. In this architecture, RNA sequences can be grouped into "islands" of conserved sequence and structure identity; however, these "islands" can exhibit significant variation with one another. Different "islands" can only be aligned with one another through the identification of common structural relationships. Based on this theoretical model and given an existing RNA sequence alignment, new RNA sequences to analyze can be grouped into one of four categories. Category 1 sequences fit completely within an existing "island" and are the easiest to align. Category 2 sequences do not fit into any existing "islands" and therefore more sequences will have to be collected in order and the CRW Project Biologist will have to establish an "island" using an iterative process of manual refinement and covariation analysis. Category 3 sequences fit within an existing

"island", but include regions of "hypervariability" which no other sequence currently identified in the "island" exhibit. Category 4 sequences fit in none of the previous three categories and therefore we assume that either: 1) they are not RNA sequences of the type contained in the existing RNA sequence alignment or 2) they are on the edge of accepted sequence space for the given RNA type and other closely related sequences are not observed in nature.

I developed the Comparative Analysis Toolkit (CAT) software package with the goal of creating a vertically integrated, expert systems infrastructure for RNA comparative analysis at the CRW Project. The initial implementation of the *Curation Pipeline* at the CRW Project was an inefficient "Rube Goldberg" style compendium of UNIX and Perl scripts and standalone C++ programs which together did not sufficiently utilize the disparate data sources such as the CRW Project RNA Metadata Database to their fullest potential to automate the analysis in a biologically relevant and meaningful manner. As a result, the CRW Project biologist was left to do a significant amount of manually repetitive work and the CRW Project had no chance of scaling to analyze all RNA sequences of interest identified in Genbank while maintaining its high standards for accuracy and quality. Given that a significant amount of software engineering work was required to lay a foundation for CAT, Category 1 and Category 3 unaligned RNA sequences have been addressed first because they are easiest given the existing diverse RNA sequence alignments available at the CRW Project. Furthermore, they provide a vehicle for demonstrating the feasibility of the infrastructure in a shorter period of time, which was necessary to convince the granting agencies. Two semi-automated sequence alignment strategies were developed within CAT to address Category 1 and Category 3 sequences (excluding regions of "hypervariability"), "*Full Alignment*" and "*Find Queries*". These strategies utilized: 1) a significantly more rigorous and efficient

implementation of "*autoalign*", 2) phylogenetic relationships between sequences and other metadata maintained in the CRW RNA Metadata Database, and 3) FASTA for rough approximation of the sequence and structural identity between an unaligned RNA sequence and an RNA sequence that has already been aligned within the existing RNA sequence alignment. A novel "*evaluator*" module was developed to analyze the quality of the alignment of an RNA sequence given known sequence and structural constraints. The "*evaluator*" reports a number of quality statistics to the Biologist and directs them specifically where to focus their efforts to refine the alignment of a sequence.

Combined, these two elements of CAT significantly streamline the analysis of Category 1 and Category 3 sequences within Stages 3 and 4 of the *Curation Pipeline*. The impact of CAT over the last 18 months on the CRW Project has been significant. The number of RNA sequences in Stage 3 of the *Curation Pipeline* has increased from 10,250 to 87,600 (Table 3.4), a 750% increase. The number of RNA sequence in Stage 4 of the Curation Pipeline has increased from 26,300 to 76,100 (Table 3.4), a 190% increase. From these results, one can conclude that CAT has been successful in facilitating the expert systems approach to RNA comparative analysis, enabling the CRW Project to demonstrate a significant increase in the number of RNA sequences analyzed by comparative analysis in the last 18 months. However, an extremely large number of identified RNA sequences still remain unanalyzed, over 800,000 (Table 3.4). While it is possible that a portion of the 800,000 remaining sequences are improperly annotated, one must conclude that the methods currently developed in CAT for higher throughout RNA comparative analysis are not sufficient to address the entire problem. Once the utility of the existing RNA sequence alignment has been exhausted, the sequence analysis methodologies currently implemented in CAT break down. In particular, analyzing Category 2 sequences and establishing new "islands" of sequence space within the

existing RNA sequence alignment is still primarily done manually using iterative alignment and covariation analysis techniques, although structure-based evaluation can contribute to facilitating the process. Furthermore, distinguishing Category 2 sequences from Category 4 sequences is also done manually. Creating, software tools to address these challenges within the expert systems framework is an active area for future development in CAT.

A number of promising research areas exist for extending CAT and the expert systems approach to RNA comparative analysis to address the analysis of Category 2 sequences. One potential method is to use data clustering techniques to attempt to pre-sort Category 2 sequences into clusters which have maximal sequence and structural identity with one another, irrespective of the existing RNA sequence alignment. Agglomerative hierarchical clustering is one technique used in the analysis of cDNA microarray data to identify patterns of gene co-expression based on a distance function[157]. We can define the distance between any two unaligned RNA sequences to be the product of their identity and overlap as computed by heuristic pairwise sequence alignment algorithms such as FASTA or BLAST. Next, unaligned RNA sequences can be clustered according to their computed distance from one another. When a cluster is identified where all sequences within that cluster are within a specific threshold distance from one another, the CRW Project biologist can manually use covariation analysis, manual alignment techniques and the "*evaluator*" to align a few sequences of that cluster as "seeds" by maximizing structure conformance to other "islands" already identified in the existing RNA sequence alignment. Finally, "*autoalign*" can be used to align the rest of the sequences identified as part of the cluster using the seed sequences that were manually aligned.

Another promising method of improving CAT with respect to the comparative analysis of Category 2 sequences is alignment through the identification of conserved RNA structural characteristics using programs such as RNAMOT[130, 131], RNAMotif[132] or ERPIN[133, 158] and libraries of structural descriptors developed from the existing RNA sequence alignment. Structural descriptors can have arbitrary complexity and can include base pairs, helices, sequence-structure motifs as well as conserved sequences. Given that the existing RNA sequence alignment already identifies "islands" of sequence identity and that we know the different islands share conserved structural constraints, it is possible to develop a library of generic descriptors where each descriptor represents a series of conserved structural constraints. Furthermore, it is possible to use patterns of sequence conservation within "islands" to create different concrete descriptors for a given generic description. Figure 3.37 is a schematic depiction of a hypothetical library of structural descriptors for a given RNA sequence alignment. To align a given Category 2 RNA sequence, each generic descriptor can be used to attempt to lock a portion of that sequence into a set of columns within the alignment. The remaining portions of the given RNA sequence which can not be aligned with any descriptor in the library are left unaligned. The alignment of a given Category 2 sequence establishes a new "island" within the existing RNA sequence alignment. As more sequences are introduced into the "island", regions which could not be addressed using the library of structural descriptors can be addressed using manual alignment and covariation analysis.

A third area for potential improvement in CAT is the "*evaluator*." Given that a large number of conserved, higher order sequence structure motifs have been identified, structural evaluation can be extended beyond simply considering known secondary structural constraints. Conserved tetraloops, E and E-like loops, lone pair tri-loops and

other sequence structure motifs can easily be mapped onto the in-memory RNA sequence alignment within CAT. Finally, the "*evaluator*" can be trained to physically manipulate the alignment of a given RNA sequence to maximize conformance to known structural constraints. A simple implementation of this idea would be to first use "*Full Alignment*" and "*Find Queries*" to rough align from a sequence perspective a Category 2 sequence using the most closely related sequence that can be identified within the existing RNA sequence alignment, regardless of the actual identity computed by FASTA. Next, the evaluator would use an iterative methodology of refining the alignment of specific nucleotides (given a limited library of potential rules governing the changes it can make) and testing the improvement against known structural constraints and rigorous covariation analysis calculations for each improvement.

Chapter 4: Overall Summary, Perspectives and Future Directions

4.A Overall Summary and Perspectives

Once it was understood that an RNA could form structure through intermolecular interactions which included DNA style secondary structure helices as well as non-canonical helices, one of the most active areas in molecular biology research has been how to predict the secondary structure for an RNA from its sequence. In the introduction to this thesis, I discuss two predominant methods for RNA secondary structure prediction: a physical chemical approach based on the experimental determination of energetic parameters and a knowledge-based approach which deduces structural relationship through the comparison of many different but homologous RNA sequences spanning the entire Tree of Life. In the knowledge-based approach, sequence and structural biases observed are subsequently fed back into the analysis as constraints reduce the complexity of the problem and facilitate the analysis of more sequences from a comparative perspective. These constraints, discussed in Section 3.B.3 and 3.B.4 can be considered "fundamentals" of RNA structure. In contrast, in the physical chemical approach, new experiments must be conduced to characterize different structural arrangements. Many times, structural arrangements predominately observed in nature such as multistem loops are difficult to study experimentally.

Programs such as Mfold[44] and the Vienna Package RNAfold[39] use a dynamic programming algorithm to predict the minimum energy secondary structure for a given sequence based on the nearest-neighbor energetics[31]. The nearest-neighbor model postulated that the free energy of a helix was a combination of the free energy of initiation and the free energy of elongation[27]. Thermodynamic parameters were measured from melting experiments using oligoribonucleotides. These parameters are the

base of the free energy function used in Mfold and RNAfold. In 1999, it was claimed that Mfold 3.1 had reached an average prediction accuracy of 73% of known base pairs[44]. In Chapter 2, I evaluated the accuracy of free-energy minimization as implemented in Mfold 3.1 for predicting RNA secondary structure from a single sequence.

This evaluation was conducted using the largest set of comparatively predicted RNA secondary structures available at the time. This set of structures included all three Ribosomal RNAs (5S, 16S and 23S) and tRNA, and included sequences as short as 72 nucleotides in length and as long as 5,461 nucleotides in length, spanning the entire Tree of Life (Section 2.C.1). The most important results of this evaluation were that: 1) the prediction accuracy of Mfold 3.1 was not significantly improved over Mfold 2.3 when one considered a large RNA comparative structure database better balanced between longer and shorter RNA sequences; 2) the accuracy of Mfold decreases significantly when one considers only comparatively predicted base pairs with a large RNA Contact Distance.

Many comparative base pairs with a large RNA Contact Distance are involved in complicated loop structures such as multistem loops and some of them are pseudoknotted, which the dynamic programming algorithm in Mfold is not capable of considering. However, failure of Mfold 3.1 to accurately predict comparative base-pairs with large RNA Contact Distance when pseudoknotted base pairs are excluded indicates that the underlying thermodynamic model for RNA secondary structure is incomplete. In particular, at the time that Mfold 3.1 was available, few reliably determined thermodynamic parameters were available for multistem loops.

In Mfold 3.1, conserved sequence-biases in multstem loops were identified from a set of comparatively predicted RNase P's[44], these biases were used in the *efn2* step which would re-rank predicted secondary structures when favorable multstem loops were

identified. The inclusion of knowledge or constraints in Mfold in this manner was not unprecedented. Tetraloops with specific sequences were first given free energy "bonuses" in Mfold 2.3. Since my evaluation, the free energy minimization algorithms have been modified to include more constraints which have further improved RNA secondary structure prediction accuracy (Section 2.F). Furthermore, new classes of algorithms including probabilistic and dynamic programming based are being developed to predict RNA secondary structure while incorporating known comparatively determined constraints, and are an intense are of research in bioinformatics and computational biology.

The results of the evaluation of Mfold combined with the number of new approaches to using comparatively predicted data were justification for the second major emphasis of this dissertation, scaling the successful methods for RNA comparative analysis developed since the late 1970s[49-51] for the large increase in the number of RNA sequences available in public repositories such as Genbank[129]. Significant advances in sequencing technology have resulted in an exponential growth in the number of sequences in Genbank since the late 1990's. The number of Ribosomal RNA sequences available has been growing at a rate as fast as Genbank, primarily due to the interest in the Ribosomal RNA for phylogenetic analysis. In Chapter 3, I discuss the Comparative Analysis Toolkit (CAT) which I have developed to streamline and scale the manual RNA comparative analysis process at the CRW Project.

The concept of comparative analysis can be defined abstractly if one considers the strategy of studying a black box system by comparing many different environmental samples of the output of that system using a constant basis. If one considers the process by which an RNA folds into its secondary and ultimately tertiary structure to be a black box system, then we can analyze RNA folding using comparative analysis. The concept

of phylogenetic conservation of function though structural conservation as postulated by Woese and Fox in 1975 was the fundamental basis for comparing different RNA sequences. The first RNA secondary structure to be predicted from a comparative basis using this principle was the 5S rRNA[49]. Subsequently, the minimal 16S and 23S rRNA secondary structure models were postulated and initially verified with chemical modification and enzymatic digestion experiments[50, 51]. Numerous other RNAs have been studies from a comparative perspective as detailed in Section 3.B.

Once the minimal 16S and 23S rRNA secondary structure models were postulated, as more RNA sequences spanning the entire Tree of Life became available, they were manually integrated into the analysis and the comparative structure models subsequently refined in light of the new sequences[52, 55]. Statistical methods were developed to systematically deduce common patterns of variation from an RNA multiple sequence alignment[52, 53]. These patterns of variation were subsequently utilized to help bring more divergent sequences in to the alignments in an iterative fashion. As a result of the systematic analysis of large, phylogenetically diverse sequence set, many RNA sequence structure motifs were discovered. The determination of the high-resolution X-ray crystal structures for LSU and SSU rRNA in 2000 subsequently validated the comparative analysis paradigm as applied to rRNA[25].

Because of the success of the initial databases of 16S and 23S rRNA secondary structure models published on the web[56-58, 159] in the search for fundamental sequence-structure motifs to characterize RNA structure, the CRW Project was started to collect, organize, analyze and disseminate RNA comparative analysis data for different RNA of interest to the scientific community[62]. The CRW Project's methods for analyzing RNA sequences were based on the manual techniques developed in the determination of the comparative structure models for 16S and 23S rRNA and the

expertise of the primary investigator, Dr. Robin Gutell, who was directly involved in developing many of these techniques. These techniques were incorporated into a systematic *Curation Pipeline*.

The manually intensive *Curation Pipeline* at the CRW Project was immediately overwhelmed when the new RNA sequences of interest identified in Genbank exceeded 79,250 in July 2003. By March 2007, over 970,000 rRNA, tRNA and Group I and II intron seqeuences had been identified in Genbank (Table 3.1). Recognizing the high level of accuracy and precision from the manual RNA sequence analysis process compared with commonly available sequence-based alignment programs such as ClustalW (Figure 3.36)[160], a hypothesis about the biological architecture of an RNA sequence alignment, and the potential contributions to RNA structure knowledge which could be gained from analyzing these RNA sequences with comparative techniques I developed the Comparative Analysis Toolkit (CAT). The primary goal of CAT was to streamline and address time-consuming and repetitive tasks which act as bottlenecks in CRW Project *Curation Pipeline*.

An analysis of the CRW Project *Curation Pipeline* revealed three bottlenecks which could be addressed by CAT. The first bottleneck involved the "*autoalign*" program initially developed by the CRW Project to align RNA sequences using "template" sequences from existing RNA sequence alignments. The "*autoalign*" algorithm was completely re-developed to align sequences more robustly map alignment results in the existing RNA sequence alignments. The second bottleneck involved the manual selection of suitable "template" sequences for a given unaligned RNA sequence. The "*Full Alignment*" and "*Find Queries*" modules were developed to systematically select existing RNA sequences to use as "templates" guided by metadata collected in the CRW Project RNA Metadata Database and rough sequence-based identity computation using FASTA.

The third bottleneck was the manual refinement, clean-up and verification of "*autoalign*" results. A novel "*evaluator*" module was developed to automatically identify errors in the alignment of an RNA sequence given known sequence and structure constraints, compute metadata data information about the alignment of an RNA sequence and update that information automatically in the CRW Project RNA Metadata Database.

The CAT software package has been under development since late 2003/early 2004. In the last 18 months, CAT has been used extensively within the CRW Project and has contributed to significant gains in the number of RNA sequences analyzed by the CRW Project (Section 3.E). One can conclude that CAT has contributed to streamlining and improving the *Curation Pipeline* at the CRW Project; however, a significantly large number of RNA sequences still remain to be analyzed. Two new methods for addressing these sequences were proposed and will be of primary emphasis in the future development of CAT. One method involves more robust techniques for sorting and grouping unaligned RNA sequences into blocks, minimizing the amount of manual sequence alignment effort required to solve a large block of sequences with "*autoalign*" and a second method involves leveraging the advances in the development of probabilistic methods for RNA sequence alignment combined with techniques for identifying conserved higher-order sequence structure motifs. However, both these techniques have an important caveat: they will only align an RNA sequence in regions where conformance to known patterns of conservation and variation can be identified. For "hypervariable" regions, manual alignment and covaration analysis will still be the primary method for identifying common secondary structure, if there is any. Furthermore, these "hypervariable" regions of the RNA sequence alignment will remain unsolved until enough sequences have been collect such that comparative analysis is possible.

CAT was not intended to replace but assist the biologist. The critical thinking skills and intuition of the biologist are intangibles difficult to completely automate, and CAT as it exists today is far from able to completely solve the RNA sequence alignment problem; however, it addresses bottlenecks in the CRW Project *Curation Pipeline* in a systematic manner without sacrificing accuracy for high-throughput. Furthermore, CAT provides a foundation for future research through incorporating: 1) the extensive libraries of conserved sequence-structure motifs and 2) careful application of the probabilistic techniques for RNA sequence alignment which have become available in the intervening time in which CAT was developed.

In summary, techniques for predicting RNA secondary structure from its sequence are an active area of research within the overall domain of the RNA Folding Problem. Given the knowledge of high-resolution X-ray crystal structures for many different RNAs and corresponding development of comprehensive RNA comparative data sets over the last 25 years, many "fundamentals" of RNA structure or sequence-structure motifs have discovered. The fundamental argument of this dissertation is that these "fundamentals" of RNA structure will provide the necessary constraints to accurately and reliably predict RNA structure from sequence in a knowledge-based manner.

Most of the community has accepted that the nearest-neighbor thermodynamic model does not sufficiently described RNA structure from a first principles perspective and comparatively determined sequence-structure motifs have been included in the thermodynamic model to improve RNA structure prediction accuracy. Using advanced software engineering techniques, tools can be developed to streamline RNA Comparative Analysis and create an expert system environment in which a Biologist can still utilize their manual intuition to continue to decipher the "fundamentals" of RNA structure. The

CAT software package as discussed in this dissertation provides a foundation for that expert system, based on techniques developed over the last 25 years to apply comparative analysis to Ribosomal RNA sequence analysis. Finally, as a corollary, in Section 4.B I discuss the software engineering framework for the next version of CAT, which will move it one step closer to being a complete expert system, and in Section 4.C I introduce a conceptual framework for utilizing a database system directly in the RNA Comparative Analysis *Curation Pipeline*.

4.B COROLLARY: APPLICATION OF CLIENT/SERVER PROGRAMMING TECHNIQUES IN THE DESIGN OF THE FUTURE VERSIONS OF CAT

The current version of CAT, 0.2.25 discussed in Chapter 3 has been designed to directly address bottlenecks RNA comparative analysis *Curation Pipeline* of the CRW Project. As was emphasized in Section 3.C.1, the software architecture inside CAT was designed to be modular, flexible and scaleable. While a significant amount of development work has gone into CAT since 2004, the program is still not optimal. Many software engineering improvements can be made to transform CAT-0.2.25 into a more powerful RNA comparative analysis toolkit which can be distributed throughout the general scientific community. Potential areas of improvement in CAT-0.2.25 include: 1) extending CAT to support more input and output formats for RNA sequences, RNA sequence alignments and RNA structural relationships. Currently CAT only supports the AE2 and FASTA formats for RNA sequence alignments and the BPSEQ format[62] for RNA structural relationships; 2) Build the RNA comparative analysis functionality within CAT including "*autoalign*" (Section 3.C.2) and the "*evaluator*" (Section 3.C.3) directly into a multiple sequence alignment editor. This functionality would allow users to interactively view results within the context of the alignment without have to export results from CAT into the alignment editor; 3) incorporate the covariation analysis

routines used by the CRW Project; 4) combine an enhanced alignment editor with integrated phylogenetic and structural navigation techniques; 5) Map additional RNA structural information into CAT including three-dimensional crystal structure data and higher order sequence-structure motifs libraries. In the rest of this section I present a proposal for the architecture of a new version of CAT which would include these concepts.

The new architecture for CATis classified as a service-oriented distributed architecture and is summarized in Figure 4.1. First, CAT will be split into two major components, a *CATServer* and a *CATGUI*. The *CATServer* will be a significant revision of the existing CAT application discussed in Chapter 3. *CATServer* will still load RNA sequence alignments into memory and will include the existing RNA comparative analysis routines such as "*autoalign*" (Section 3.C.2) and the "*evaluator*" (Section 3.C.3). The *CATServer* will still have the capabilities of a standalone command-line based application, but will also provide a programmatic interface by which other applications can directly "hook" in and utilize the functionality through remote procedure calls.

The *CATGUI* will be built as a separate component. *CATGUI* can be executed on the same computer as the *CATServer*, or it can be located on a different computer. This separation is necessary for scalability. Our goal is to have the ability to manipulate extremely large RNA sequence alignments efficiently through a graphical user interface. By separating the graphical user interface in to a separate process, the user has to opportunity to execute the *CATServer* on a computer that has enough memory and compute resources to facilitate the manipulation and analysis of large RNA sequence alignments while still possessing the ability to interact with those RNA sequence alignments from their laptop via the graphical user interface.

4.B.1 *CATServer* Engine

The architecture of CAT-0.2.25 is discussed extensively in Section 3.C.1 and is summarized in Figure 3.12. CAT is architected to load into memory and analyze sequence alignments with more than 10^5 sequences and up to $5x10^3$ columns. An alignment of this size would have a memory footprint of roughly 3.7GB which is close to the limit for a single process on most 32-bit computers. Since 64-bit computers were not very prevalent late 2003/early 2004 when this architecture was first conceived and given the known size of the RNA sequences sets that have been identified by the CRW Project at that time, it was decided to save time and implement the C++ data structures for 32-bit platforms (Section 3.C.1.1). On hindsight, this was an incredibly short sighted "Y2K" style decision on my part. Therefore the *CATServer* will have the ability to execute on 32-bit or 64-bit platforms. With this re-development, *CATServer* will be capable of loading multiple RNA sequence alignments in excess of 10^6 sequences.

Additionally, CAT-0.2.25 is extremely limited in the number of input and output formats it supports. In particular, the AE2 alignment format, created in the late 1980's when AE2 was initially developed (Section 3.B.2) is the only supported input format for RNA sequence alignments. This restriction is in place due to the fact that the only alignment editor used by the CRW Project is the AE2 alignment editor. Because we will be developing a new alignment editor, *CATGUI* (Section 4.B.2) we have the opportunity to eliminate the AE2 format and replace it with a more robust RNA sequence alignment format that promotes interoperability with the rest of the RNA community, RNAml[161].

Using RNAml format, we can embed structural and phylogenetic relationships directly within the RNA sequence alignment file. Furthermore, AE2-formatted RNA sequence alignments have a significant amount of bloat due to the direct embedding of additional annotation symbols beyond gaps. With a new RNA sequence alignment format

based on RNAml and embedded structural information, we can remove all of the extra embedded annotation from the alignment file, significantly reducing file size and the in-memory footprint for an RNA sequence alignment while eliminating the need for the aligned sequence mapping step in the current "*autoalign*" (Section 3.C.2). Using the embedded structural information, the *CATGUI* will now project structural annotation directly into the alignment view, and will support navigation of the alignment using embedded phylogenetic and structural relationships (Section 4.B.2).

CATServer will include a number of RNA comparative analysis enhancements including the direct integration of covaration analysis methods used by the CRW Project. Users will be able to invoke covariation analysis routines directly from the *CATGUI* and interactively view the results (Section 4.B.2). Furthermore, the *CATServer* will have the ability to map three-dimensional crystallographic data and higher order sequence structure motifs directly into the in-memory alignment and utilize the data in RNA comparative analysis. The power of displaying higher order sequence-structure motifs directly with secondary structure and other comparative data has been demonstrated with the RNAMap (Gandhi et al, unpublished data) interactive structure viewer prototype available at the CRW Project website[62]. If the strategies suggested in Section 3.E for improving RNA sequence alignment are successful, they will be included in the *CATServer*.

From a software infrastructure perspective, the most important new functionality in *CATServer* will be the addition of a remote interface for linking *CATServer* with *CATGUI* and incorporating *CATServer* as a component within other software toolkits developed by the scientific community. This programmatic interface will be based on XML over Java Remote Method Invocation (RMI). Any RNA comparative analysis routine that can be performed on an RNA sequence alignment loaded in-memory can be

invoked remotely, programmatically. Further more, any "chunk" of an in-memory sequence alignment for display in an alignment editor can be directly requested programmatically, a feature that will be heavily utilized by the *CATGUI*.

4.B.2 CAT Graphical User Interface with Integrated Alignment Editor (*CATGUI*)

The architecture for future versions of CAT is to be service-oriented and distributed (Figure 4.1). In Section 4.B.1 I discussed the "*CATServer*". In this section, I discuss the *CATGUI* which will be a platform-independent, extensible, graphical user interface (GUI) application that is mouse, menu and toolbar driven where users can: 1) interactively browse and edit alignments or two-dimensional structure data loaded in-memory in the *CATServer*, 2) execute comparative analysis routines on RNA sequence alignments loaded in-memory in the *CATServer* and view their results annotated directly within the alignment editor view, and 3) provide phylogenetic and structural methods to navigate large RNA sequence alignments efficiently with biological purpose.

To speed application development, the open-source Eclipse framework (http://www.eclipse.org) will be utilized as the basic GUI toolkit eliminating the need to develop menus, tool bars, dialog boxes, mouse support, drag and drop, a basic text editor, a context-dependent help framework, integrated software updating, and more. The Eclipse framework is platform-independent, and although it is Java-based, Eclipse outperforms other Java Swing-based GUI implementations due to the superior performance of the Standard Widget Toolkit (SWT). SWT's primary advantage is its ability to utilize native operating system supplied widgets (e.g., dialog boxes, menus, windows, etc.) in a platform-independent manner. The Eclipse framework is based on a modular "plug-in" architecture. With this architectural approach, functionality can easily be added, removed and modified to augment and customize the functionality of the

framework. Our domain specific functionality will be implemented as Eclipse framework plug-ins.

4.B.2.1 The Distributed Alignment Editor Plug-in

The Distributed Alignment Editor Plug-in is the means to viewing large alignments in the *CATGUI*. Figure 4.2 provides an illustration of how this plug-in works. Only a small fragment of the alignment will be loaded in the memory space of the CATGUI at any point in time, and the portion of that fragment visible to the user is function of the individual user's screen resolution and font settings. Network bandwidth will not be an issue. Consider that if the dimensions of the fragment were 100 columns X 100 rows at 2 bytes per grid coordinate, we would have a 20 KB fragment if transferred in binary format. By comparison, most broadband home Internet connections are 128 KB/sec or faster. Furthermore, if the *CATGUI* and *CATServer* are co-located on the same computer, bandwidth issues are not a problem as inter-process communication can occur via named pipes. The Distributed Alignment Editor Plug-in will contain all the basic features found in most alignment editors available today including multiple editing modes, mouse or keyboard driven, configurable, differential coloring schemes, and masks. In addition to basic features, the distributed alignment editor plug-in will contain features to facilitate working with very large sequence alignments in a distributed manner. The features are:

Views: Views are subsets of the complete alignment that display a limited range of nucleotide positions for a limited set of sequences. Edits to the alignment introduced within any view window are immediately reflected in all other open views. For example, in a type 1 tRNA alignment, a user may want to study the D-loop of the cloverleaf structure in Bacterial vs. Archaea tRNAs. The user can open two separate views of the tRNA alignment loaded in the *CATServer* (Figure 4.3), one focused on the Archaea

sequences and a second view focused on the Bacterial sequences. The number of simultaneous views a user can have open is limited by the amount of memory installed in the computer where they are executing the *CATGUI*.

Annotation: Results of RNA comparative analysis routines executed on the *CATServer* will be displayed and annotated directly in the alignment editor view. For example, "hotspots" detected by the *MismatchAccumulationAnalyzer* in the "evaluator" module (Section 3.C.3.3) will be directly annotated in the alignment view. Figure 4.3 is an example of how a "hotspot" could be directly annotated in the alignment editor window. Base pair and nucleotide conservation can be indicated directly in the alignment editor viewport using differential coloring

Structure Browser: The purpose of the "Structure Browser" (Figure 4.3) is to facilitate movement across large numbers of columns in the alignment using structural relationships. This structure-based navigation feature utilizes the integration of structural relationships directly into the alignment. A dedicated "Structure Browser" window displays the mapping of the currently-selected sequence to the common secondary structure for that alignment. Selecting a particular structural element (e.g., a base-pair) scrolls the current alignment view to the column containing the first nucleotide of the structural element. For example, if a given alignment had a base-pair relationship between column 1000 and column 5000, the current alignment view would immediately be scrolled to column 1000. To display the region of the alignment for the other half of the base-pair (column 5000), the user has the option to split the current view or open a new view. Selecting a different sequence in the alignment would cause the "Structure Browser" window to be updated with the mapping of the newly-selected sequence on the common secondary structure.

Phylogeny Browser: Another way to efficiently navigate large sequence alignments is to use the evolutionary relationships in the data. Phylogenetic relationships between sequence will either be directly included in the RNAml version of the RNA sequence alignment, or they can be fetched from the CRW RNA Metadata Database. The "Phylogeny Browser" window (see Figure 4.3) contains a tree view of the taxonomy. Users can filter the view so that all nodes are present, or just nodes containing sequences. As the user clicks on different nodes of the tree, the alignment view is scrolled to the first sequence that is part of that node.

4.C COROLLARY: A PROPOSAL FOR AN ADVANCED DATA MANAGEMENT FRAMEWORK FOR RNA COMPARATIVE ANALYSIS

In revolutionizing biological sequence analysis, bioinformatics and computational biology have imposed a general workflow paradigm. While the specific algorithms for analyzing different types of biological sequence and structure data are a source of constant research and development, the general workflow has remained the stagnant. In particular: 1) Computational biology and bioinformatics software tools expect biological sequence datasets to be stored in flat files located on the file system of a computer server; 2) Databases are primarily used to hold metadata, and there is little if any direct integration of sequence, structure and evolutionary information in the database system; 3) Analysis programs operate on the sequence data in the flat files in-memory. In some cases, the results of the analysis are stored in a database, but in many cases the analysis is re-executed when it is required.

The biggest limitation of this workflow paradigm will be scalability. As public sequence databases such as Genbank grow exponentially, in-memory approaches to biological sequence analysis will have difficulty scaling to handle extremely large datasets without expensive hardware systems. This dissertation has focused on scaling

issues related RNA comparative analysis from an expert system perspective. These scaling issues result from an ever increasing number of RNA sequences available in Genbank. One method for addressing these issues is through leveraging advances in databases and the introduction of distributed, service-oriented software architectures based on scalable and componentized middleware technologies and distributed computing. In order to do this, one must fundamentally change their view of what is primary data and what is a computationally derived entity. Currently biological sequences and experimentally determined structures are considered primary data which is persisted in a database with a service-oriented interface (Genbank and the PDB respectively).

In the particular case of RNA comparative analysis, one can envision that the sequence alignment should not be considered a computationally derived entity. Rather, the alignment itself and the complex evolutionary and structural relationships that are associated with sequences in the alignment are primary data dimensions (Figure 4.4), and not computationally derived entities. Therefore, the sequence alignment and the complex evolutionary and structural relationship should be directly persisted in a database. The data model should maintain the grid like structure of the alignment. Structural relationships should be modeled as n-ary relationships between columns of the alignment grid and phylogenetic relationships as hierarchical relationships between rows of the alignment grid. Phylogenetic relationships impose a hierarchical ordering on the sequences in the alignment, while structural relationships link different, potentially distant columns in the alignment. The presence or absence of specific structural relationships can also vary with phylogeny. (e.g., a particular base-pair relationship between two columns in the alignment that is only observed for Archaea species).

A database system capable of storing the alignment and indexing it by its phylogenetic and structural relationships provides the opportunity to replace a significant

amount of software infrastructure with database queries. The database, which is already highly optimized for execution planning, parallel processing, resource management, and data storage retrieval provides the software infrastructure. Indexes are defined in terms of the embedded phylogenetic and structural relationships to improve the efficiency of the system. A few sample queries which could be executed directly in the database with a series of queries include: 1) single nucleotide frequency queries for specific branches of the phylogenetic tree (e.g., nucleotide frequency for column 100 for the Archaea); 2) structural composition and frequency queries for simple base-pairs or aggregates such as helices or stem loops (e.g., the base-pair between columns 9 and 25 in Bacteria is 95% G-C); 3) conservation of a given structural element in specific branches of the phylogenetic tree (e.g., a particular helix is five base-pairs in the Archaea, but shortens to three base-pairs in the Bacteria and Eukaryota).

One potential argument against this approach is that complex computations such as covariation analysis cannot be represented as simple database queries. As a result, external (to the database) software tools will still be required for these complex computations. Furthermore, these programs will have to retrieve the sequence alignment from the database into their own memory space in a client/server manner which could be potentially disastrous for performance (Note: the client/server architecture proposed for the CAT in Section 4.B does not suffer from this issue because the comparative analysis algorithms are implemented in the *CATEngine* component, which is where the RNA sequence alignments are loaded in-memory) as a huge amount of data must be marshaled from the database to the analysis program. These concerns are valid, and any suitable database system for this infrastructure must facilitate the development of arbitrary modules which are implemented in procedural programming languages such as Java and

can be executed within the database process. Complicated computational algorithms would then be implemented as database modules.

4.C.1 Prototype Implementation Based on a Hierarchical Database

Working in collaboration with Dr. Philip Cannata, of Sun Microsystems we built a prototype database implementation using Sun One Directory Server (SODS). SODS is a hierarchical database. In a hierarchical database, a given database schema is modeled as a series of "nodes". Every node has a specific set of indexable attributes, one parent node and zero or more child nodes. The database access protocol is Lightweight Directory Access Protocol (LDAP), analogous to SQL for relational database. One downside to using a hierarchical database is that LDAP is not nearly as flexible and rich as SQL for writing database queries. With Sun One Directory Server (SODS), we have the ability to build our own database plug-ins. Database plug-ins are compiled C libraries that execute in their own thread inside the NDS process. Data does not have to be communicated out of the database process prior to being utilized in a computation plug-in. Plug-ins are instantiated via LDAP queries to the database. Any arguments required for a plug-in to execute are specified in the LDAP query.

Figure 4.5 is a cartoon schematic depicting how an RNA sequence alignment is persisted in a hierarchical database such as SODS. The general database topology follows the taxonometric relationships defined for the entire Tree of Life. The database contains two different kinds of nodes: taxonomy nodes and alignment row nodes. Taxonomy Nodes can contain other Taxonomy Nodes and Alignment Row Nodes; Alignment Row Nodes are leaf nodes. Each Alignment Row Node contains both a set of metadata attributes and a set of column attributes. Figure 4.5 shows how any particular row of the alignment is mapped in the database. The number of Alignment Grid Cell attributes in each Alignment Row Node is equivalent to the number of columns in the alignment, and

for a given alignment every Alignment Row Node has the same number of column attributes. The value of each Alignment Grid Cell is the corresponding value from the alignment (e.g., A, G, C, U, -). In our initial prototype, column attributes were simple character types containing the character at that column in the alignment. Metadata attributes provide more general information about a row including: availability (public or private), sequence length, cellular location (nucleus, mitochondrion or chloroplast), etc.

Structural relationships which are n-ary relationships between columns (Figure 4.5) are implemented as n-ary relationships between column attributes for any alignment row node. A base-pair would be implemented as a relationship between two specific column attributes of the alignment row node, the columns which contain the two nucleotides in the base-pair (Figure 4.5). Logical aggregations of these relationships map to specific structural elements. For example, a RNA secondary structure helix is two or more consecutive base-pairs. In the database, this is represented as relationships between consecutive columns.

To establish the feasibility of using a hierarchical database system for persisting sequence alignments, we have built a prototype using the Sun One Directory Server (SODS). The SODS comes from the same code base as the open-source Netscape Directory Server (NDS), which we plan to use for this project. The prototype was developed by making a series of modifications to SODS to allow nodes with greater than 104 attributes. With this prototype and a SSU rRNA alignment which contained 43,200 sequences, 12,227 columns, and spanned the entire Tree of Life, we performed a series of simulations to measure the performance for a series of the most common queries. The types of queries we measured for performance broke down into three major classes:

Class 1: *Looking down any single column of the alignment, determine the distribution of nucleotides for that column*. An example result would be: column 100 is 95% G, 4%C and 1% others.

Class 2: *Looking down any two-related columns, determine the distribution of nucleotides for that pair of columns*. An example result would be: column 100 – column 200 is 95% G-C and 5% C-G. This type of query would be utilized to quickly determine frequencies of base-pairs.

Class 3: *Looking down any set of eight consecutive columns, determine the distribution of nucleotide for that set of columns*. The concept is similar to a Class 1 query; however, queries of this style would be utilized for determining both the average length and nucleotide composition of secondary structure helices.

In addition to organizing our test queries into classes, we also had to devise a mechanism for measuring the impact of the depth and breadth of the phylogenetic tree on any particular query. Because we are using a hierarchical database, the height and depth of the hierarchy directly impacts the query performance. The reference hierarchy we utilized came from the NCBI Taxonomy Database. Since our test alignment contains sequences spanning the entire Tree of Life, one can understand that executing a Class 2 query across all sequences in the alignment would be equivalent to executing a Class 2 query across the full depth of the database. In contrast, executing a Class 2 query across only the Crenarchaeota, limits the query to only a portion of the database. To account for the depth and breadth issues in a standardized manner, the queries were randomly started at specific heights in the hierarchy and traversed the entire subtree below their starting point. A query could start at a height of 1, 5, 10, 15, 25, 35, or 40. A height of 1 is equivalent to starting at the root of the hierarchy while a height of 40 was equivalent to starting at node which was 40 levels down from the root of the hierarchy.

Simulations which were thirty minutes in length were carried out for each class of query, with a load of 100 consecutive users, starting at different depths in the database. The simulations were carried out on a Sun Fire V880 with eight processors and 65 GB RAM. Of the 65 GB of RAM, 50 GB were allocated to the database cache, and approximately 30,000 entries were present in the cache. The results are summarized in Table 4.1. The system supported an average of 101 Class 1 queries per second, 77 Class 2 queries per second and 31 Class 3 queries per second. Overall, the system supported approximately 70 queries per second. Due to the hierarchical manner in which the data is stored in the database, the performance is significantly faster when a query begins deeper in the hierarchy. These numbers demonstrate that our prototype system can adequately support 100 simultaneous users in the most common query scenarios. Even though the database cache was 50 GB for this prototype, we are confident that the amount of cache required will drop by as much as a factor of 10 or more once other optimizations are made to the database system

	5S rRNA	16S rRNA	23S rRNA	tRNA[1]	Total
Structure Models	90	496	256	569	1,411
Total AGCU Nucleotides	10777	724475	712575	42283	1,490,110
Total Nucleotides	10819	736412	714723	43189	1505143
Total Comparative Pairings[2]	3107	191994	178958	11796	385854
Average Sequence Length	120	1485	2792	76	-
Average Pairings/Structure[2]	35	387	699	21	-
Phylogenetic Distribution					
Archaea	12	23	17	76	128
Bacteria	28	195	75	155	453
Eucarya					
Nuclear	45	133	52	207	437
Chloroplast	4	33	31	131	199
Mitochondrion	1	112	81	-	194

[1] Includes only G:C, A:U and G:U base-pairings predicted with comparative analysis

[2] Only Type I tRNAs are considered

Table 2.1: RNA comparative structure database

Relative composition of the RNA comparatively predicted structure database used for the evaluation of Mfold 3.1.

	5S rRNA		16S rRNA			23S rRNA			tRNA	
	M[1]	C[2]	P_1[3]	M	C	P_2[4]	M	C	M[5]	C
Sequences	309	90	56	22	496	72	5	256	484	569
Average Accuracy	78 ± 23	71 ± 24	46 ± 17	51 ± 16	41 ± 13	44 ± 11	57 ± 14	41 ± 13	83 ± 22	69 ± 24
Previous Study[6]					45 ± 16			43 ± 12		
High/Low[7]		98/0	81/0		77/5	74/19		74/1		100/0
Median		81			41			41		70
Distributions										
≤20% Acc[8]		4	9		4	1		6		2
≥60% Acc[9]		77	25		9	6		5		60
20%<Acc<60%[10]		19	66		86	93		89		39

[1] All sequences from the Mathews et al study (M) were folded with Mfold 3.1 using a window size of 0, percent suboptimality 20%, maximum number of suboptimals of 750 and efn2 re-evaluation and re-ordering

[2] All sequences in the the current study (C) were folded with Mfold 3.1 using a windows size of 1, percent suboptimality of 5% and efn2 re-evaluation and re-ordering

[3] All sequences in the previous Gutell Lab study on 16S rRNA (P_1) were folded with Mfold 2.3 using a windows size of 10 and no efn2 re-evaluation and re-ordering

[4] All sequences in the previous Gutell Lab study on 23S rRNA (P2) were folded with Mfold 2.3 using a windows size of 20 and no efn2 re-evaluation and re-ordering

[5] Bases modified in tRNA that are subsequently unable to fit into an A form heilx were constrainted to be single-stranded

[6] Average prediction accuracy using Mfold 3.1 using just the subset of sequences considered in the previouse Gutell Lab studies (**P_1**, **P_2**)

[7] Accuracy scores for the best and worst predicted structures in each group

[8] Percentage of predicted structures with an accuracy of 20% or less

[9] Percentage of predicted structures with an accuracy of 60% or higher

[10] Percentage of predicted structures with an accuracy between 20% and 60%

Table 2.2: Average accuracy of the optimal RNA secondary structure predicted with Mfold 3.1

Results are reported for 5S, 16S and 23S Ribosomal RNA (rRNA) and Transfer RNA (tRNA). All values are percentages unless otherwise indicated. C, Current Study; P1, Mfold 2.3 evaluation by the Gutell Lab using 16S rRNA [40]; P2, Mfold 2.3 evaluation by the Gutell Lab using 23S rRNA [41]; M Mfold 3.1 evaluation by Mathews et al. [44]. Accuracies from all previous studies are for folding complete sequences. A discussion on how secondary structure prediction accuracies are computed can be found in Section 2.G.1. All averages are computed as per sequence averages as discussed in Section 2.G.2.

	5S rRNA	16S rRNA		23S rRNA		tRNA
	C	P_1	C	P_2	C	C
Archaea	79	68	62	59	58	73
Bacteria	62	56	49	53	49	74
Eucarya (n)[1]	75	30	34	41	42	61
Eucarya (c)[1]	67	48	46	39	39	73
Eucarya (m)[1]		31	30	38	30	
Eucarya (m)[1,2]			31			
Eucarya (m)[1,3]			33			

[1] (n), Nuclear-encoded sequences; (c) Chloroplast-encoded sequences; (m) Mitochondrial-encoded sequences

[2] Based on comparative models with 100 or more canonical base pairs only

[3] Based on comparative models with 300 or more canonical base pairs only

Table 2.3: Average accuracy of the optimal RNA secondary structure predicted with Mfold 3.1 grouped by phylogenetic classification

All values shown are percentages unless otherwise indicated. C, Current Study; P1, Mfold 2.3 evaluation by the Gutell Lab using 16S rRNA [40]; P2, Mfold 2.3 evaluation by the Gutell Lab using 23S rRNA [41]. A discussion on how secondary structure prediction accuracies are computed can be found in Section 2.C.4.5. All averages are computed as Per Sequence averages as discussed in Section 2.G.2

	Previous (P_1, P_2)	Current
16S rRNA		
Eukaryotic Mitochondrion		
Zea Mays (X00794)	17	30
Ascaris summ (X54253)	17	13
Caenorhabditis elegans (X54252)	23	24
Eukaryotic Nuclear		
Hexamita sp. (Z17224)	27	29
Giardia muris (X65063)	22	29
Giardia ardeae (G17210)	30	33
Giardia intestinalis (X52949)	10	23
Encephalitozoon cuniculi (X98467)	18	21
Vairimorpha necatrix (Y00266, M24612)	28	25
Babesia bigmina (X59064)	20	19
23S rRNA		
Eukaryotic Chloroplast		
Astasia longa (X14386)	19	23
Eukaryotic Mitochondrion		
Caenorhabditis elegans (X54252)	30	31
Gallus gallus (X52392)	28	25
Sacccharomyces cerevisiae (J01527)	27	20
Zea mays (K01868	24	29
Eukaryotic Nuclear		
Euglena gracilis (X53361)	23	21
Giardia intestinalis (X52949)	24	33

Table 2.4: Accuracy of the optimal RNA secondary structure predicted with Mfold 2.3 and Mfold 3.1 for specific 16S and 23S Ribosomal RNA (rRNA) sequences.

For all RNA sequences in this table, the optimal RNA secondary structure prediction by Mfold 2.3 was 30% or less. All values are percentages unless otherwise indicated. P1, Mfold 2.3 evaluation by the Gutell Lab using 16S rRNA [40]; P2, Mfold 2.3 evaluation by the Gutell Lab using 23S rRNA [41]. Genbank [129] accession numbers are listed in parentheses for each sequence. A discussion on how secondary structure prediction accuracies are computed can be found in Section 2.G.1.

	16S rRNA		23S rRNA	
RNA Contact Distance	496 Structures		256 Structures	
Comparative				
Total	191,994	100%	178,958	100%
2-100	145,058	76%	134,085	75%
2-50	121,170	63%	106,534	60%
51-100	23,888	12%	27,551	15%
101+	46,936	24%	44,873	25%
101-500	43,004	22%	37,121	21%
501+	3,932	2%	7,752	4%
Predicted Correctly		**%Comp**[1]		**%Comp**[1]
Total	81,934	43%	77,888	44%
2-100	75,763	52%	67,130	50%
2-50	64,651	53%	54,898	52%
51-100	11,202	47%	12,232	44%
101+	6,171	13%	10,758	24%
101-500	5,978	14%	9,441	25%
501+	193	5%	1,317	17%
Average Per Sequence Accuracy	**C**	**P_1**	**C**	**P_2**
2-100	50%	55%	47%	53%
2-50	52%		49%	
51-100	44%		40%	
101-200	22%	15%	26%	35%
201-300	10%	14%	22%	21%
301-400	9%	13%	13%	10%
401-500	4%	12%	16%	13%
501+	4%		14%	

[1] The percentage of comparatively predicted base pairs observed in the specified RNA Contact Distance range.

Table 2.5: Accuracy of individual base pairs predicted with Mfold 3.1 as a function of RNA Contact Distance.

All base pairs from the set of 16S and 23S Ribosomal RNA (rRNA) comparative structure models are grouped by RNA Contact Distance and their accuracy is determined collectively. P1, Mfold 2.3 evaluation by the Gutell Lab using 16S rRNA[40]; P2, Mfold 2.3 evaluation by the Gutell Lab using 23S rRNA [41]. RNA Contact Distance is defined as the number of nucleotides intervening between the 5' and 3' halves of a base pair. A discussion on how Per Sequence averages are computed can be found in Section 2.G.2.

	16S rRNA		23S rRNA	
RNA Contact Distance	496 Structures		256 Structures	
Predicted with Mfold 3.1				
Total	223,957	100%	218,908	100%
2-100	150,886	67%	137,780	63%
2-50	123,708	55%	109,078	50%
51-100	27,178	12%	28,702	13%
101+	73,071	33%	81,128	37%
101-500	43,498	19%	44,139	20%
501+	29,573	13%	36,989	17%
Predicted Correctly		**%Mfold[1]**		**%Mfold[1]**
Total	81,934	37%	77,888	36%
2-100	75,763	50%	67,130	49%
2-50	64,651	52%	54,898	50%
51-100	11,202	41%	12,232	43%
101+	6,171	8%	10,758	13%
101-500	5,978	14%	9,441	21%
501+	193	0.7%	1,317	4%

1 The percentage of Mfold 3.1 predicted base pairs observed in the specified RNA Contact Distance range.

Table 2.6: Mfold 3.1 Predicted base pairs grouped by RNA Contact Distance

All base pairs from the set of optimal 16S and 23S Ribosomal RNA (rRNA) structure models predicted with Mfold 3.1 are grouped by RNA Contact Distance and their accuracy is determined collectively. RNA Contact Distance is defined as the number of nucleotides intervening between the 5' and 3' halves of a base pair.

	Overall	Archaea	Bacteria	Eucarya		
				Chloroplast	Mitochondrion	Nuclear
Comparative	191,994	10,211	83,385	13,406	29,979	55,013
Optimal Correct[1]	81,934	6,376	41,032	6,105	9,459	18,962
Suboptimal Correct[2]	137,000	8,570	65,177	10,032	21,201	32,020
Optimal Incorrect[1]	142,023	4,758	49,563	8,603	27,617	51,482
Suboptimal Incorrect[2]	2,372,305	101,253	947,197	161,397	472,614	689,844
Optimal Accuracy[1]	41%	62%	49%	46%	30%	34%
Suboptimal Accuracy[2]	71%	84%	78%	75%	71%	59%
Average Improvement[3]	30%	21%	29%	30%	41%	24%
Best Prediction[4]	92%	91%	89%	92%	92%	90%
Max Improvement[5]	68%	35%	54%	53%	68%	48%
Min Improvement[6]	10%	10%	12%	12%	14%	11%

[1] Total number of comparative base pairs observed in Mfold 3.1 optimal secondary structure predictions for the 496 comparatively prediction 16S rRNA secondary structure models.

[2] Total number of comparative base pairs observed in the Mfold 3.1 optimal + 749 suboptimal secondary structure predictions for the 496 comparatively predicted 16S rRNA secondary structure models.

[3] Average improvement in Mfold 3.1 secondary structure prediction accuracy when pooling base pairs from both the optimal and 749 suboptimal predictions.

[4] The highest Mfold 3.1 secondary structure prediction accuracy for an individual 16S rRNA sequence when pooling base pairs from both the optimal and 749 suboptimal predictions.

[5] The largest improvement in Mfold 3.1 secondary structure prediction accuracy for an individual 16S rRNA sequence when pooling base pairs from both the optimal and 749 suboptimal predictions.

[6] The smallest improvement in Mfold 3.1 secondary structure prediction accuracy for an individual 16S rRNA sequence when pooling base pairs from both the optimal and 749 suboptimal predictions.

Table 2.7: The distribution of 16S Ribosomal RNA (rRNA) comparatively predicted base pairs predicted correctly considering the optimal and 749 suboptimal secondary structure predictions from Mfold 3.1.

All 496 16S rRNA comparatively prediction secondary structure models are considered. Values are calculated by summing the number of unique base pairs encountered for each sequence with in the optimal or optimal + suboptimal population. For example, Suboptimal Correct is calculated by summing the number of unique, correctly predicted base pairs encountered in a population of 750 secondary structure predictions (1 optimal + 749 suboptimal structure predictions). A discussion on how secondary structure prediction accuracies are computed can be found in Section 2.G.1. All averages are computed as Per Sequence averages as discussed in Section 2.G.2.

	RNA Contact Distance						
	2-100 nt		**101-500 nt**		**501+ nt**		**Total**
Correct[1]	115,471	*84%*	20,069	*15%*	1,460	*1%*	137,000
Never[2]	29,587	*54%*	22,935	*42%*	2,472	*4%*	54,994
Total	145,058	*76%*	43,004	*22%*	3,932	*2%*	191,994

[1] Total number of comparative base pairs observed in the Mfold 3.1 optimal + 749 suboptimal secondary structure predictions for the 496 comparatively predicted 16S rRNA secondary structure models.

[2] Total number of comparative base pairs never observed in the Mfold 3.1 optimal + 749 suboptimal secondary structure predictions for the 496 comparatively predicted 16S rRNA secondary structure models.

Table 2.8: Frequency of comparatively predicted base pairs in Mfold 3.1 predicted secondary structures as a function of RNA Contact Distance

The distribution by RNA Contact Distance of 16S Ribosomal RNA (rRNA) comparatively predicted base pairs considering the optimal and 749 suboptimal secondary structure predictions from Mfold 3.1. The comparatively predicted base pairs are grouped into two categories: 1) observed at least once and 2) never observed. All 496 16S rRNA comparatively prediction secondary structure models are considered.

	Approximate Numbers of RNA Sequences			
	July 2003			March 2007
	Aligned[1]	Unaligned[2]	Total	Total
16S rRNA	19,000	24,000	43,000	495,400
23S rRNA	2,400	27,600	30,000	216,000
5S rRNA	1,400	100	1,500	6,600
Group I Introns	2,000	200	2,200	4,000
Group II Introns	900	750	1,650	1,400
tRNA	900	0	900	250,000
Totals	26,600	52,650	79,250	973,400

[1] A given sequence is considered "aligned" once it has passed through al

[2] A given sequence is considered "unaligned" if it has not passed

Table 3.1: The number of RNA sequences identified and analyzed by the CRW Project between July 2003 and March 2007.

Sequence counts are rounded to the nearest 50. Unaligned sequences identified are based on an annotation-based search of Genbank [129] only.

Approximate Number of RNA Sequences at Different Stages of the "*Curation Pipeline*" in July 2003				
	Stage 2	**Stage 3**	**Stage 4**	**Total**
16S rRNA	15,800	8,200	19,000	43,000
23S rRNA	26,500	1,100	2,400	30,000
5S rRNA	100	0	1,400	1,500
Group I Introns	0	200	2,000	2,200
Group II Introns	0	750	900	1,650
tRNA	300	0	600	900
Total	42,700	10,250	26,300	79,250

Table 3.2: Distribution of RNA sequences identified by the CRW Project in July 2003 throughout Stages 2, 3 and 4 of the *Curation Pipeline*

Sequence counts are rounded to the nearest 100.

	Pairwise Identity				
	<=70%	**71-80%**	**81-90%**	**91-94%**	**>=95%**
Bacteria					
16S					
Accuracy[1]	76.68%	79.82%	92.56%	97.30%	95.99%
Structural Identity[2]	94.73%	95.97%	98.93%	99.76%	99.75%
Pairs Tested	6	24	72	20	78
23S					
Accuracy	73.27%	85.60%	91.42%	95.24%	98.65%
Structural Identity	92.02%	95.85%	97.61%	99.35%	99.82%
Pairs Tested	46	24	18	20	24
Eukaryota					
16S					
Accuracy	78.44%	83.04%	90.13%	92.28%	94.44%
Structural Identity	87.17%	90.26%	97.15%	98.42%	99.34%
Pairs Tested	24	44	60	12	56
23S					
Accuracy	75.95%	84.50%	87.89%	94.25%	96.78%
Structural Identity	93.23%	94.27%	96.83%	98.20%	99.30%
Pairs Tested	6	6	18	6	6
Mitochondrion					
16S					
Accuracy		80.59%	93.41%	98.25%	99.08%
Structural Identity		95.07%	97.06%	98.53%	99.29%
Pairs Tested		14	10	6	14
23S					
Accuracy	63.51%	75.21%	97.07%	98.26%	98.65%
Structural Identity	92.74%	94.02%	99.70%	99.90%	99.80%
Pairs Tested	14	10	6	10	2

[1] Accuracy for the pairwise recursive dot-plot alignment methodology in "*autoalign*"

[2] Structural Identity is defined as the ratio of overlapping nucleotides to total columns in the pairwise alignment of two sequences (see Figure 3.5)

Table 3.3: The accuracy of the heuristic pairwise alignment algorithm in "*autoalign*"

The accuracy for the heuristic pairwise alignment algorithm is determined for pairs of 16S and 23S rRNA sequences within different phylogenetic groups. The results are separated into five different groups based on the sequence identity between the "template" and "query" sequences (Section 3.C.2.1). Structural Identity and Sequence Identity are defined in Section 3.B.5.1. The accuracy is computed by comparing the result of the pairwise alignment algorithm with the known manual alignment result.

Approximate Number of RNA Sequences at Different Stages of the "*Curation Pipeline*" in March 2007				
	Stage 2	**Stage 3**	**Stage 4**	**Total**
16S rRNA	360,400	75,000	60,000	495,400
23S rRNA	196,200	11,800	8,000	216,000
5S rRNA	2,600	0	4,000	6,600
Group I Introns	1,000	0	3,000	4,000
Group II Introns	100	800	500	1,400
tRNA	249,400	0	600	250,000
Total	809,700	87,600	76,100	973,400

Table 3.4: Distribution of RNA sequences identified by the CRW Project in March 2007 throughout Stages 2, 3 and 4 of the *Curation Pipeline*

Sequence counts are rounded to the nearest 100.

	Query Type		
Start Height	**Class 1**	**Class 2**	**Class 3**
1	14.06	11.92	6.19
5	23.7	17.98	8.34
10	34.85	23.41	13.8
15	39.69	20.67	12.1
25	155.1	116.9	23.2
35	223.7	177.6	80
40	221	174.5	75.3
Average	101.7	77.57	31.3

Table 4.1: Performance simulation results for 16S Ribosomal RNA (rRNA) sequence alignment of 43,200 sequences and 12,227 columns loaded into Sun One Directory Server (SODS) according to the database schema in Figure 4.5.

The simulation was run for 30 minutes for a load of 100 simultaneous users. Results reported are the total number of queries in each of the three classes processed over the 30 minute simulation. The Start Height represents the depth of the query across the phylogenetic tree. For example, a query with a Start Height of 1 begins at the root of the phylogenetic tree and is exhaustive. By comparison, a query with a Start Height of 40 begins much deeper within the phylogenetic tree. The Start Height for a query was assigned randomly. The different query classes are discussed in Section 4.C.1 and a brief summary is provided here. Class 1 queries are equivalent to a nucleotide frequency computation across a single column. Class 2 queries are equivalent to base pair frequency computations involving two columns. Class 3 queries involve eight columns and would be similar to determining the composition of an entire secondary structure helix.

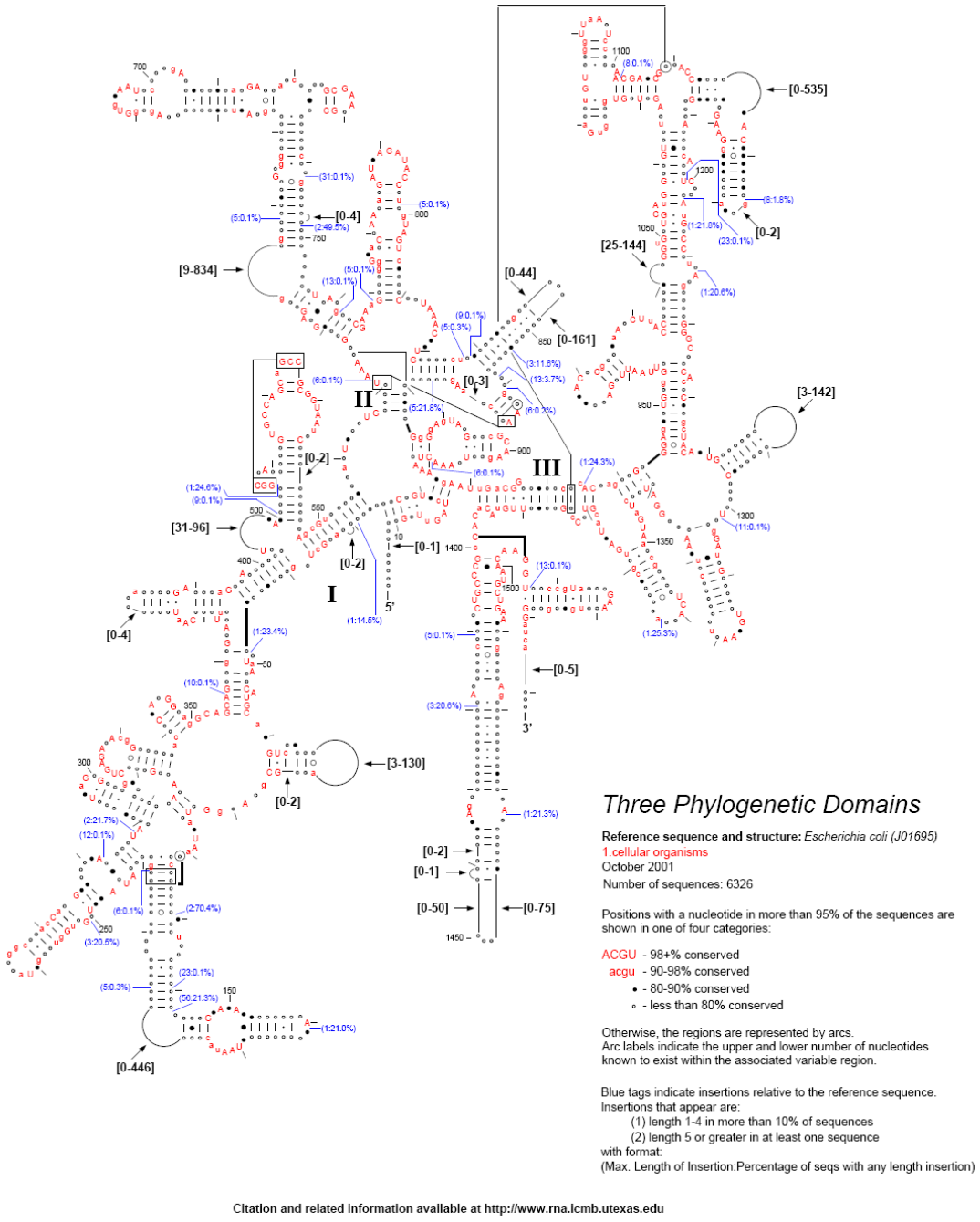


Figure 2.1: 16S Ribosomal RNA secondary structure conservation diagram from the CRW Web Site

Generated from an alignment of 6326 16S rRNA nuclear encoded sequences spanning the Tree of Life[62]. Positions in the structure model with a nucleotide in 95% or more of the sequences are depicted as a nucleotide or a circle depending on their conservation level. Variable regions are indicated with arcs which indicate their range in length.

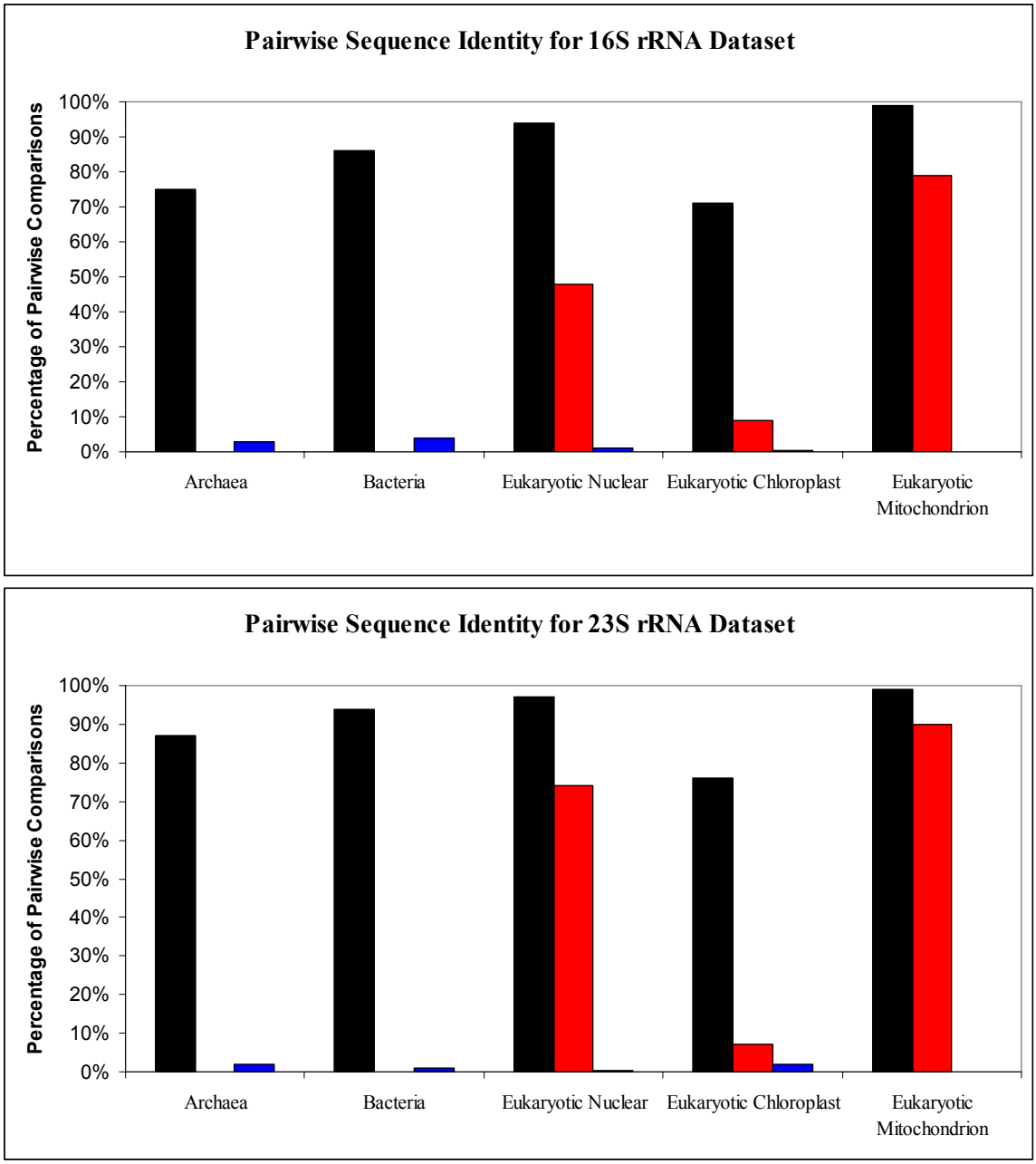


Figure 2.2: The sequence diversity in the 16S and 23S Ribosomal RNA (rRNA) data sets using pairwise identity comparisons.

Sequence pairs with less than 80% identity are represented with black bars. Sequence pairs with less than 50% identity are represented with red bars. Sequence pairs with greater than 95% identity are represented with blue bars.

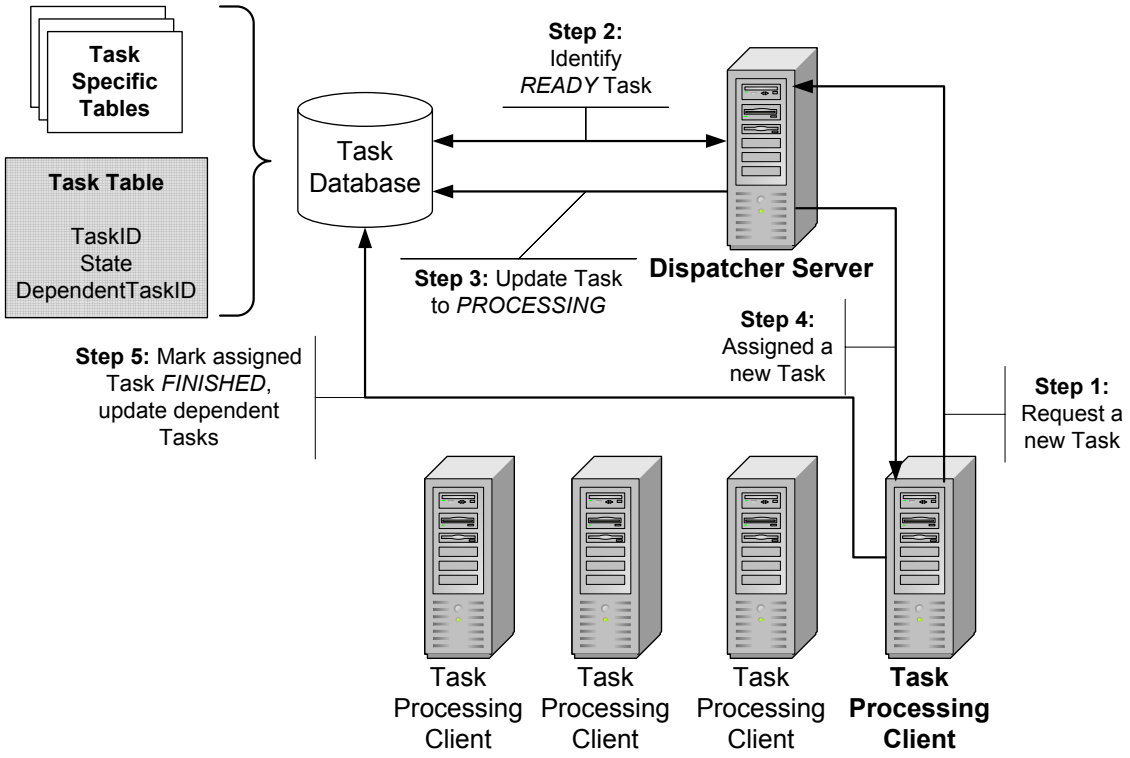


Figure 2.3: Computational setup for the evaluation of Mfold 3.1

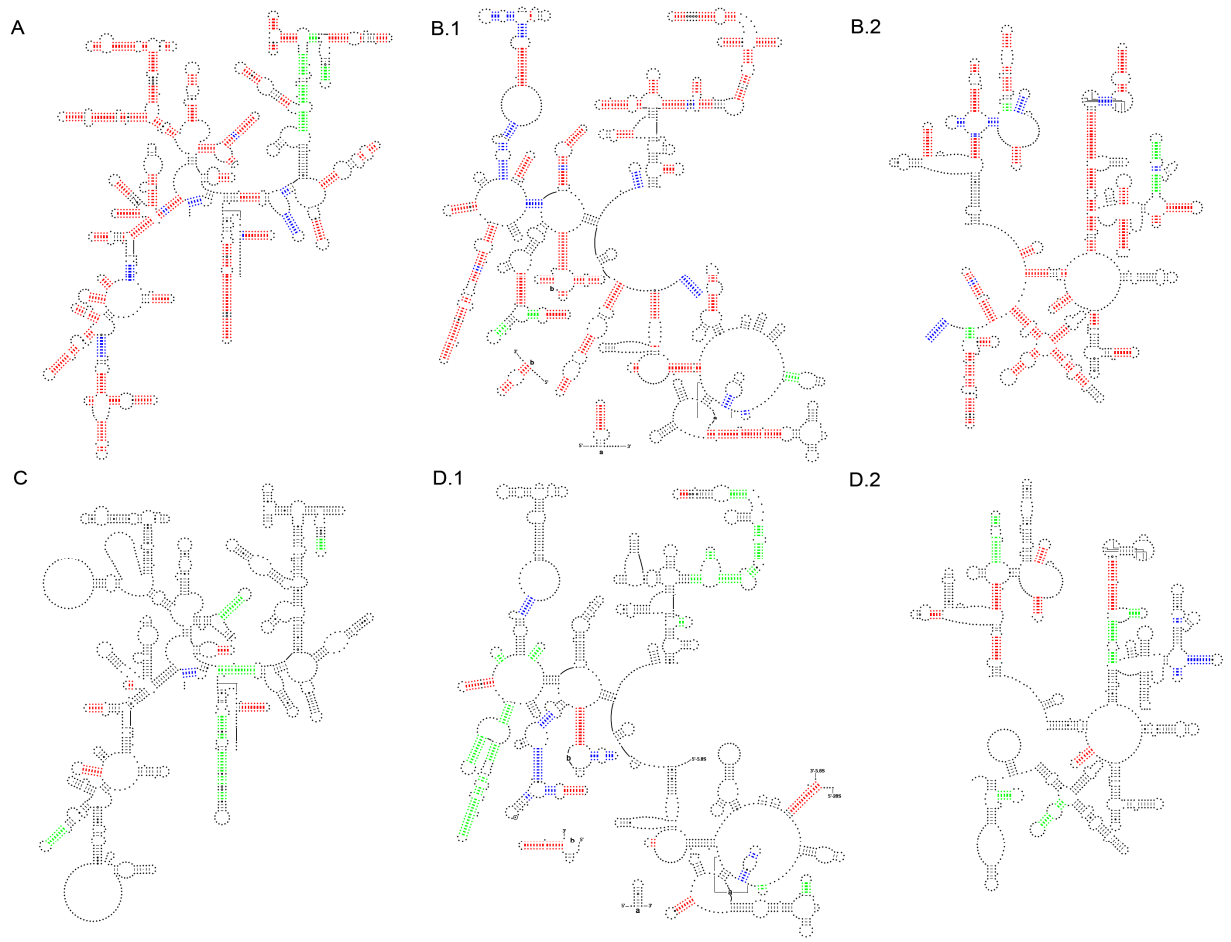


Figure 2.4: Direct comparison of Mfold 2.3 and Mfold 3.1 folding accuracies for selected 16S and 23S Ribosomal RNAs (rRNA)

Base pairs marked in red are predicted correctly by both Mfold 2.3 and Mfold 3.1. Base pairs marked in blue are predicted correctly only by Mfold 2.3, and base pairs marked in green are predicted correctly only by Mfold 3.1. Base pairs with no color designation are not predicted correctly by either version of Mfold. Only canonical base pairs (G:C, A:U and G:U). **A:** Archaea 16S rRNA *Haloferax volcanii*. **B.1:** Archaea 23S rRNA 5'half, *Thermococcus celer*. **B.2:** Archaea 23S rRNA, 3' half, *Thermococcus celer*. **C:** Eukaryotic Nuclear 16S rRNA, *Giardia intestinalis*. **D.1:** Eukaryotic Nuclear 23S rRNA, 5' half, Giardia *intestinalis*. **D.2:** Eukaryotic Nuclear 23S rRNA, 3' Half, *Giardia intestinalis*.

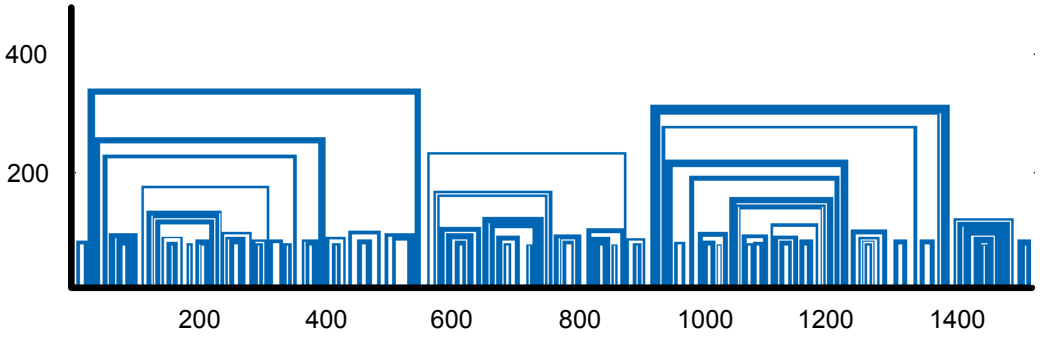


Figure 2.5: 16S Ribosomal RNA secondary structure represented as a "*histogram*"

Adapted from the from the CRW Project Web Site [62]. This plot is based on the secondary structure model for Escherichia coli which has 1542 nucleotides. The nucleotides in the sequence are represented along the X-axis. For each base pair in the secondary structure, the nucleotides are connected with a blue line. The height of the blue line is proportional to the number of nucleotides intervening between the 5' and 3' ends of the base pair.

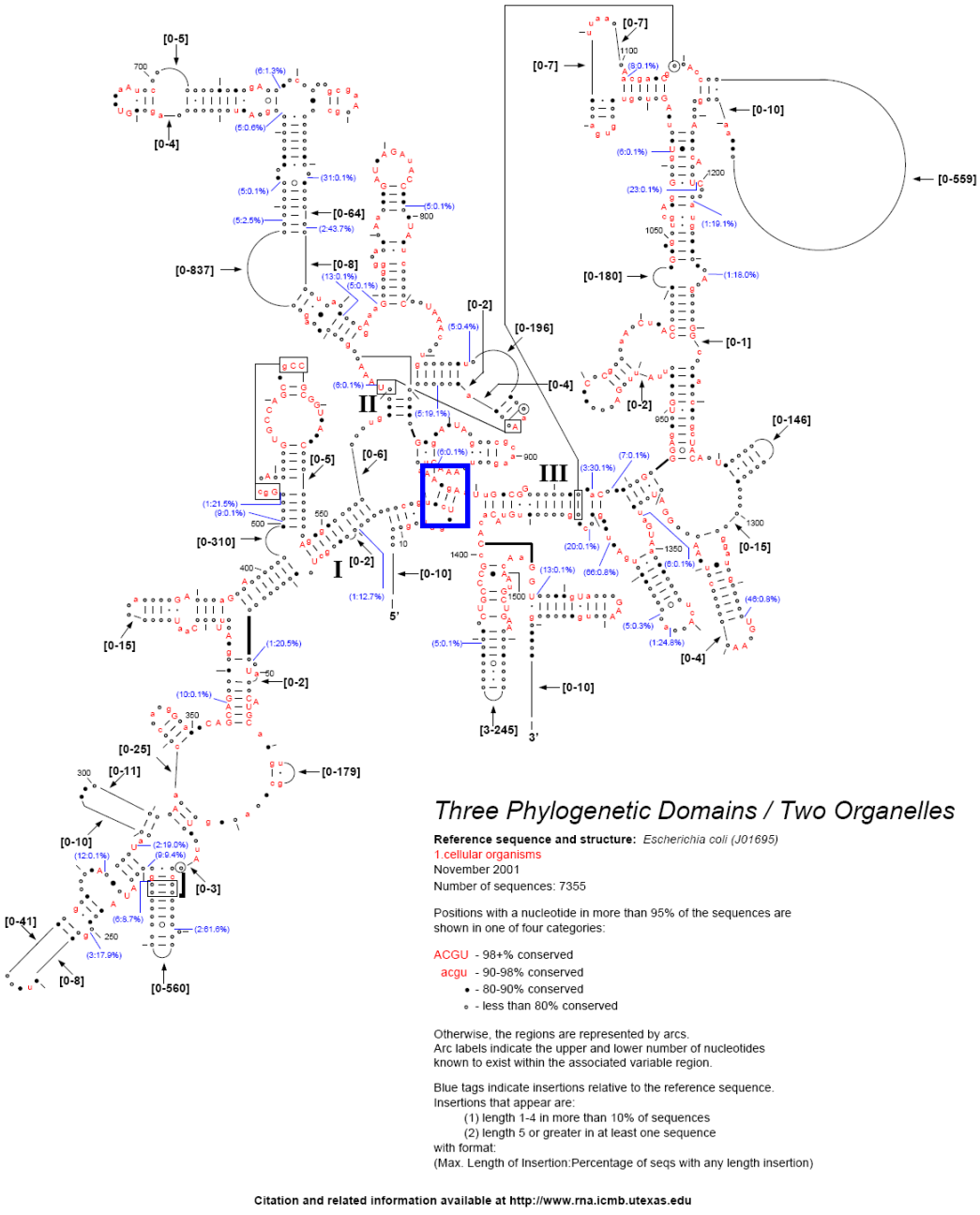


Figure 2.6: 16S Ribosomal RNA secondary structure conservation diagram

From the CRW Web Site[62] generated from an alignment of 7355 16S rRNA nuclear, chloroplast and mitochondrial encoded sequences spanning the Tree of Life. A conserved, pseudoknotted helix of long range base pairs (17:918, 18:917, and 19:916) is highlighted with a blue box.

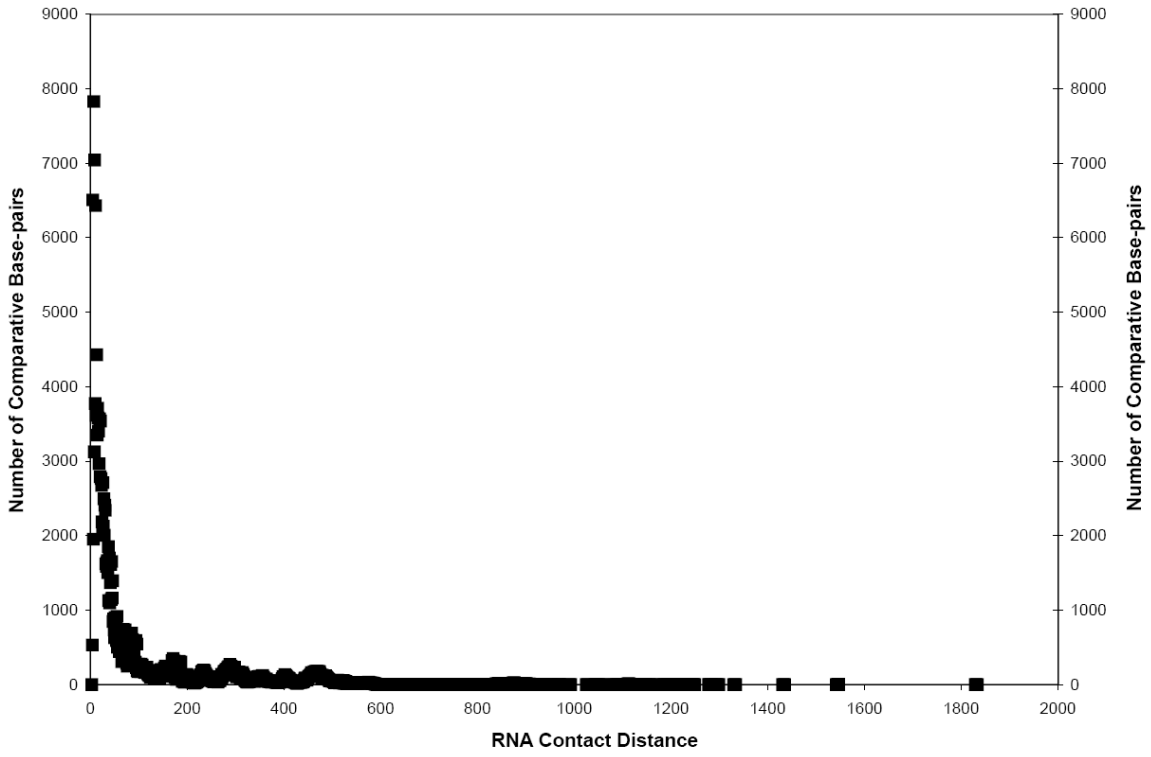


Figure 2.7: Distribution of the 191,994 comparatively predicted base pairs from 496 16S Ribosomal RNA (rRNA) secondary structure models as a function of RNA Contact Distance.

RNA Contact Distance is defined as the number of nucleotides intervening between the 5' and 3' halves of a base pair.

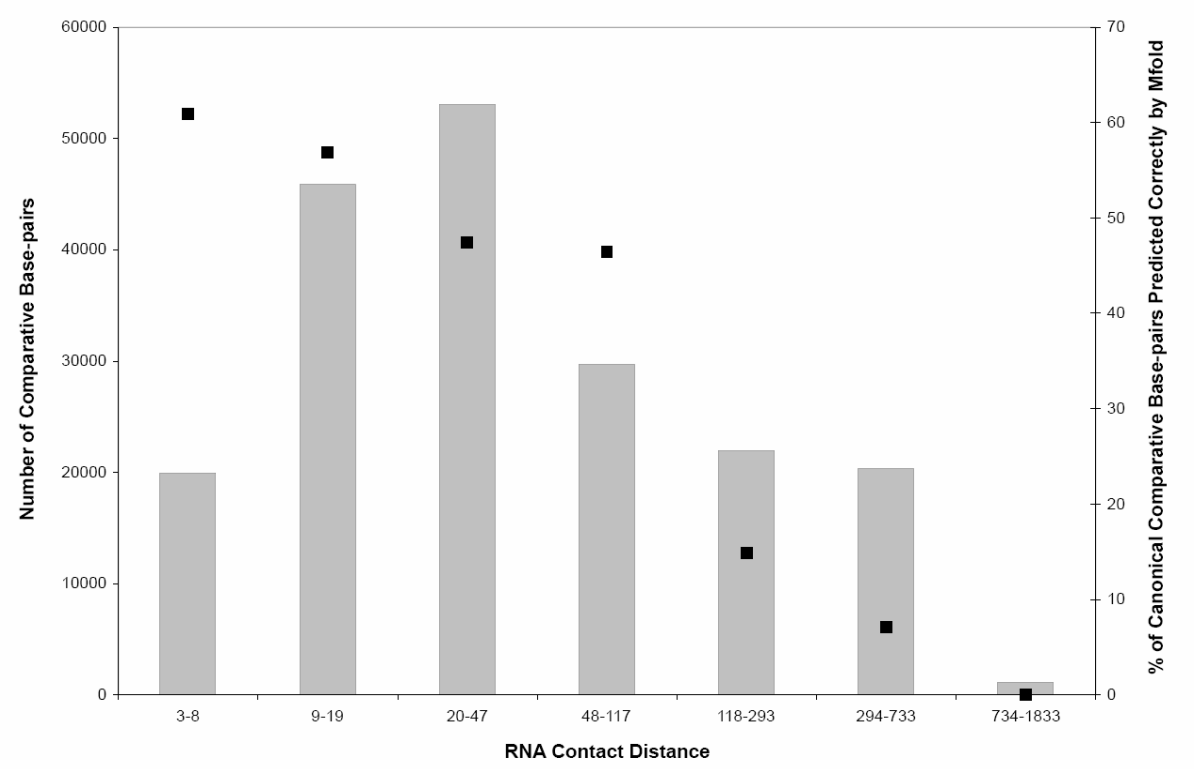


Figure 2.8: Logarithmic binning of the 191,994 comparatively predicted pairs from 496 16S Ribosomal RNA (rRNA) secondary structure models as a function of RNA Contact Distance.

The base pairs are divided into seven RNA Contact Distance bins. The average prediction accuracy (based on the optimal Mfold 3.1 secondary structure predictions for each of the 496 16S rRNA comparative structure models) for base pairs in each RNA Contact Distance bin is plotted as an individual point.

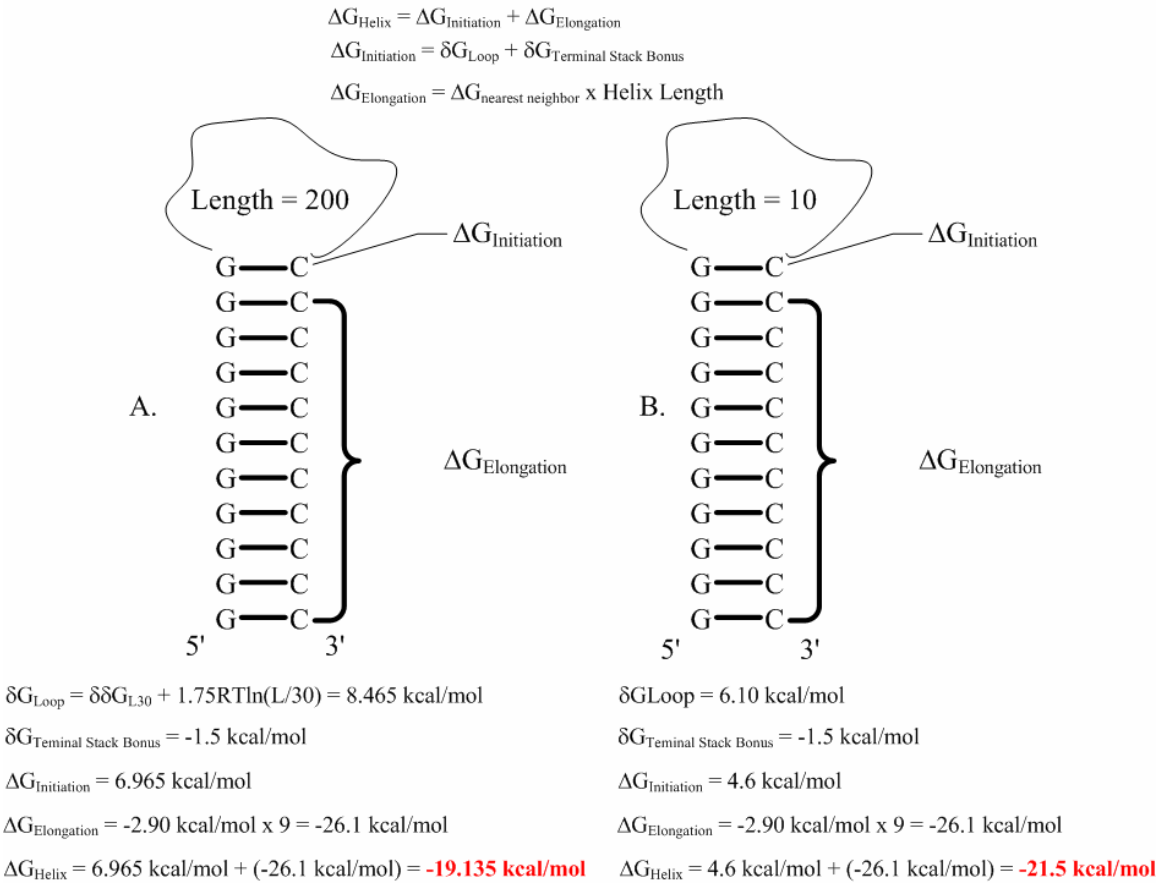


Figure 2.9: Example free energy calculation for a helix as implemented in Mfold 3.1

ΔG_{Helix} is the sum of the initiation free energy ($\Delta G_{Initiation}$) and the elongation free energy ($\Delta G_{Elongation}$). Since both helices in this example consist of consecutive GC base pairs, we only need to consider the nearest neighbor free energy for a GC duplex, -2.90 kcal/mol. For both helices, $\Delta G_{Elongation}$ is just the number of base pairs multiplied by the -2.90 kcal/mol. $\Delta G_{Initiation}$ is the sum of the destabilizing energy involved in forming the first base pair (δG_{Loop}) and the free energy from the terminal stacking base pair ($\delta G_{Terminal\ Stack\ Bonus}$). For the helix A, modified Jacobsen-Stockmeyer theory is used to extrapolate δG_{Loop} beyond length 30. For helix B, δG_{Loop} is experimentally determined.

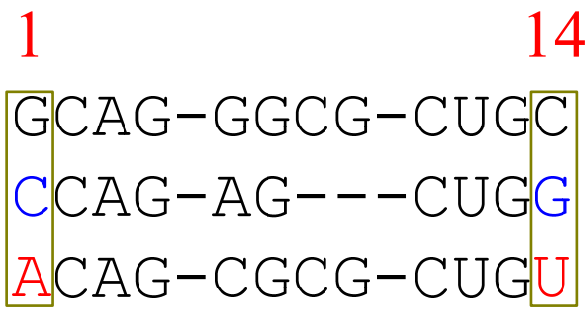


Figure 3.1: A simple example of a *positional covariation*

Column 1 and Column 14 exhibit coordinated, compensating changes to maintain a Watson-crick base pair. The first coordinated change (blue) GC => CG, and the second coordinated change (red) is CG => AU.

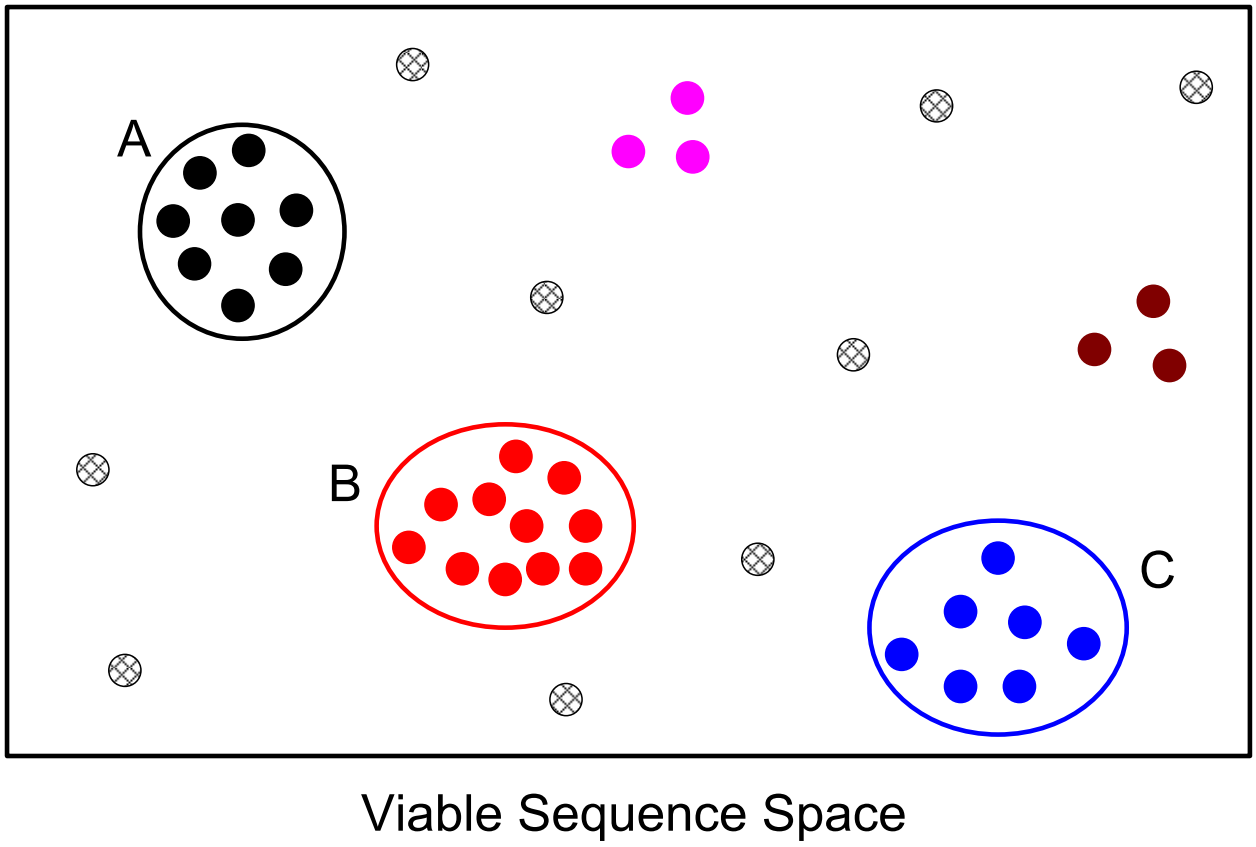


Figure 3.2: An abstract sequence space plot for a given RNA type.

Each individual circle represents a sequence sample. The agglomerated blocks of sequences (A, B and C) are considered "islands" where all sequences samples within the "island" exhibit significant sequence and structure identity with one another. "Islands" A, B and C are well-sampled and the patterns of sequence and structure conservation are established. For the other sequence samples, more sampling of sequence space is required before "islands" including those sequences can be established.

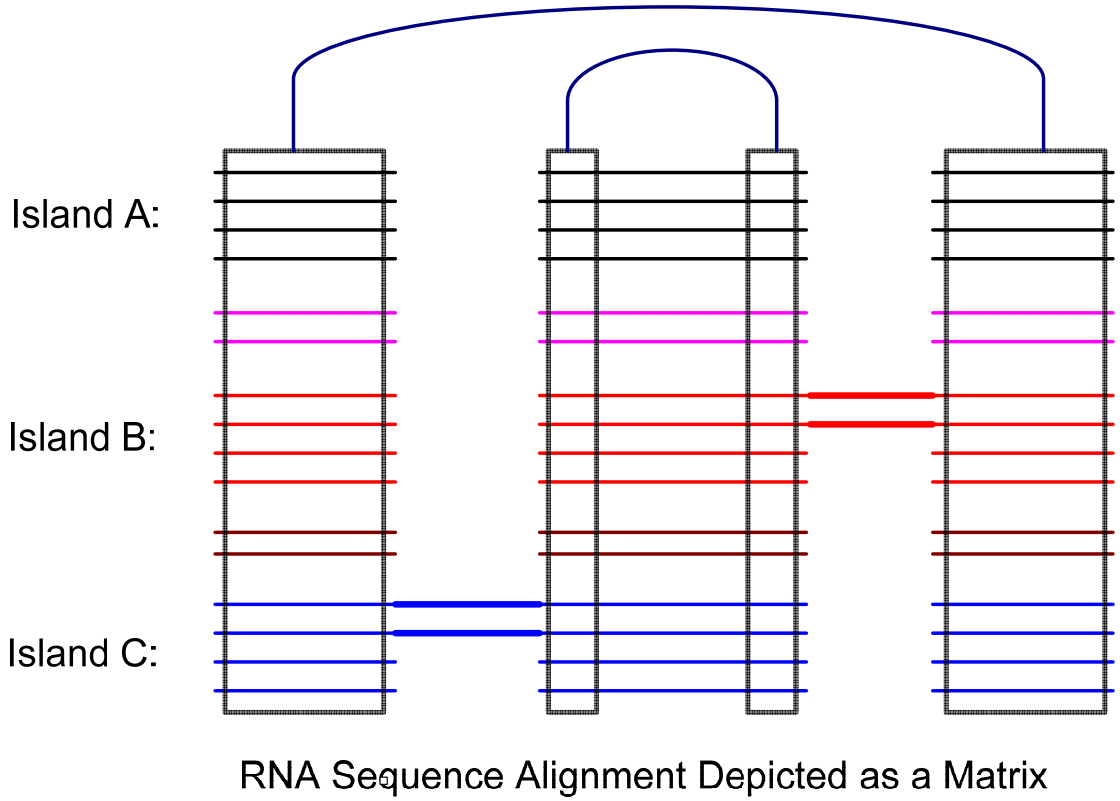


Figure 3.3: The sequence samples from Figure 3.2 arranged into an abstract RNA sequence alignment in the normal matrix view

The sequences are juxtaposed into the alignment such that individual columns of the alignment represent functionally equivalent positions between the sequences. In this example sequence samples from "Islands" A, B, and C in Figure 3.2, which do not exhibit high sequence identity with one and other are aligned based on common patterns of variation which represent common structure. The location of common structure is identified by black connected boxes. Common patterns of variation are deduced through the identification of *positional covariations*. The bold sequences in Island B and C represent regions of "hypervariability", where not all sequences within the "island" include the particular inserted sequence.

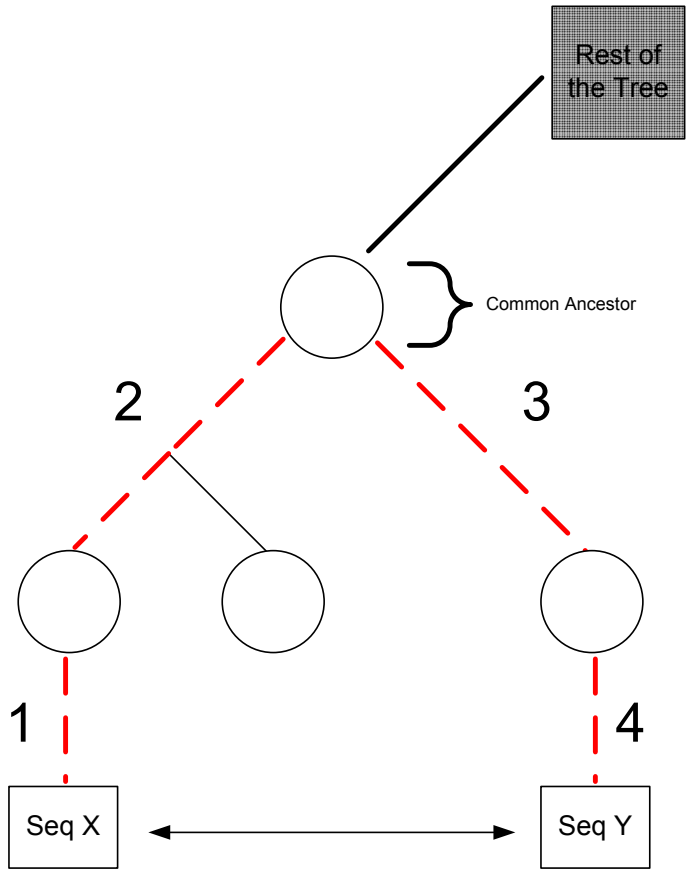


Figure 3.4: Schematic example of a "Phylogenetic Distance" computation.

Circle nodes represent parent nodes while square nodes contain the sequences X and Y for which we want to compute the "Phylogenetic Distance". Starting from sequence X and Y, we walk up the tree until we identity their common ancestor node. The "Phylogenetic Distance" is the sum of the links on the path between sequence X and Y via the common ancestor.

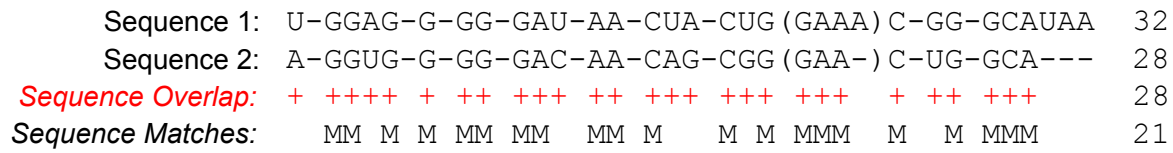


Figure 3.5: Schematic example of the "Sequence Identity" and "Structural Identity" computations

"Sequence Identity" is the ratio of the matching nucleotides (*Sequence Matches*) to overlapping nucleotides (*Sequence Overlap*) for two aligned sequences. In this example, the "Sequence Identity" is 21/28 = 75%. "Structural Identity" is measured in terms of the numerical magnitude of insertions and deletions between two aligned sequences. This magnitude is represented as the ratio of the number of columns in the alignment where the two sequences overlap (*Sequence Overlap*) to the number of columns in the alignment where either sequence has a nucleotide. In this example, "Structural Identity" is 28/32 = 88%.

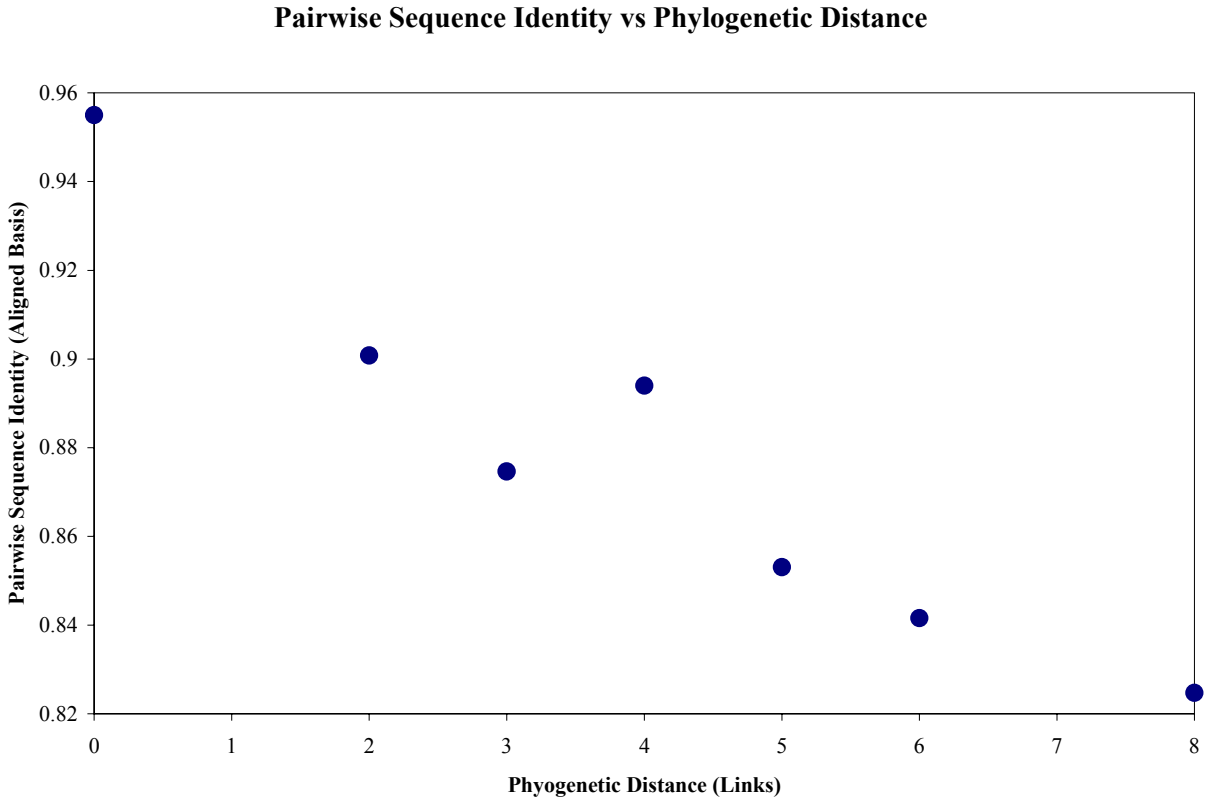


Figure 3.6: Pairwise sequence identity vs. phylogenetic distance from a 16S rRNA sequence alignment spanning the Tree of Life

An analysis of the manually generated 16S rRNA alignment spanning the entire Tree of Life from the CRW Web Site [25]is presented. This alignment contains 6326 complete sequences. All pairs of sequences within the alignment with a given Phylogenetic Distance are identified (Figure 3.4); a total 783,409 pairs. The pairwise sequence identity is computed for each pair of sequences (Figure 3.5). All computed pairwise sequence identity values are averaged together for each Phylogenetic Distance.

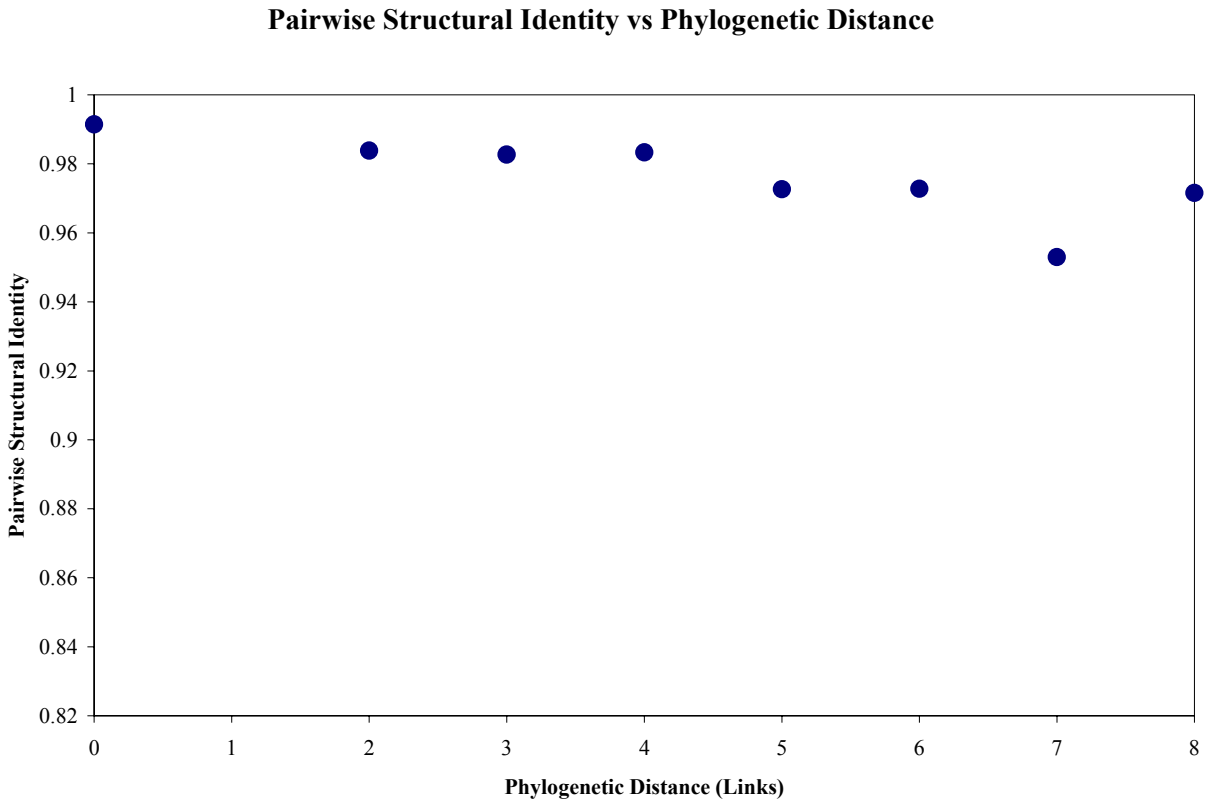


Figure 3.7: Pairwise structural identity vs. phylogenetic distance from a 16S rRNA sequence alignment spanning the Tree of Life

An analysis of the manually generated 16S rRNA alignment spanning the entire Tree of Life from the CRW Web Site is presented. This alignment contains 6326 complete sequences. All pairs of sequences within the alignment with a given Phylogenetic Distance are identified (Figure 3.4); a total of 783,409 pairs. The pairwise structural identity is computed for each pair of sequences (Figure 3.5). All computed pairwise structural identity values are averaged together for each Phylogenetic Distance.

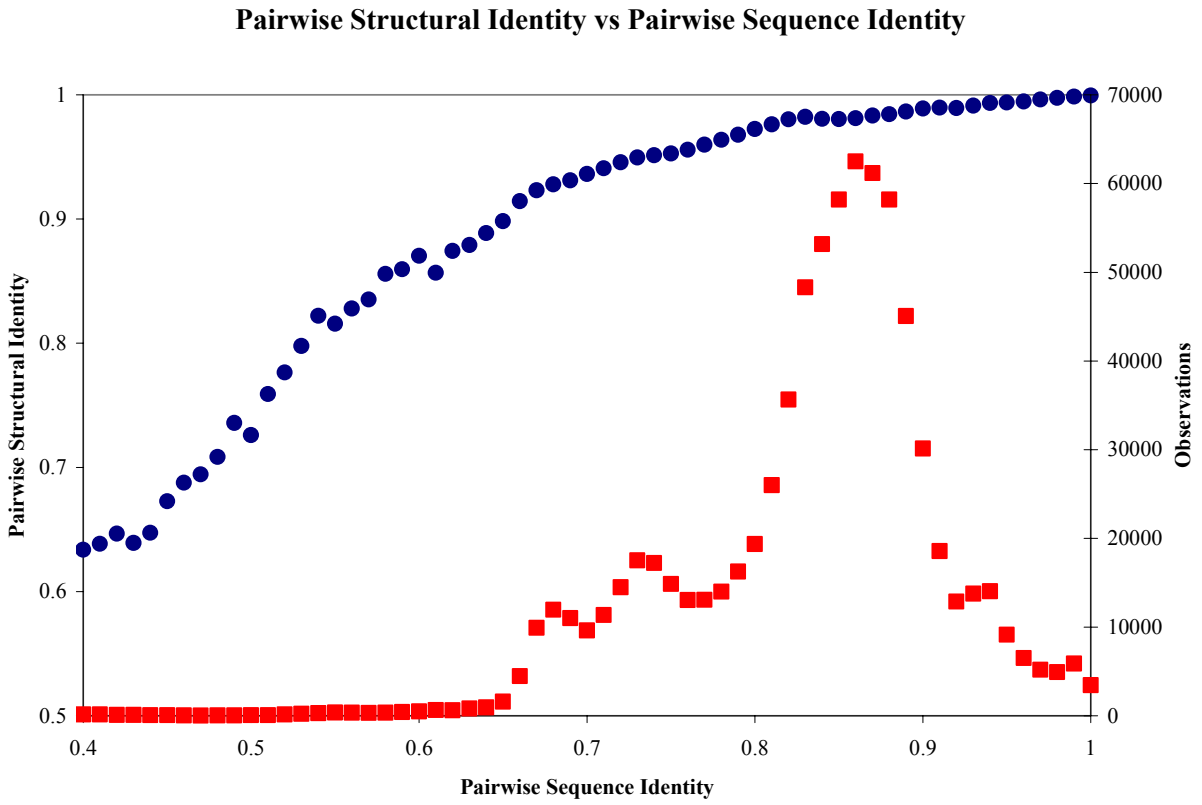


Figure 3.8: Pairwise structural identity vs. pairwise sequence identity from a 16S rRNA sequence alignment spanning the Tree of Life

An analysis of the manually generated 16S rRNA alignment spanning the entire Tree of Life from the CRW Web Site is presented. This alignment contains 6326 complete sequences. A total of 783,409 pairwise comparisons with a maximum Phylogenetic Distance of 8 are considered. The average pairwise structural identity (Figure 3.6) is plotted against the pairwise sequence identity (blue diamonds). All computed pairwise structural identity values are averaged together for each pairwise sequence identity. The number of observations at each particular sequence identity is also plotted as red squares.

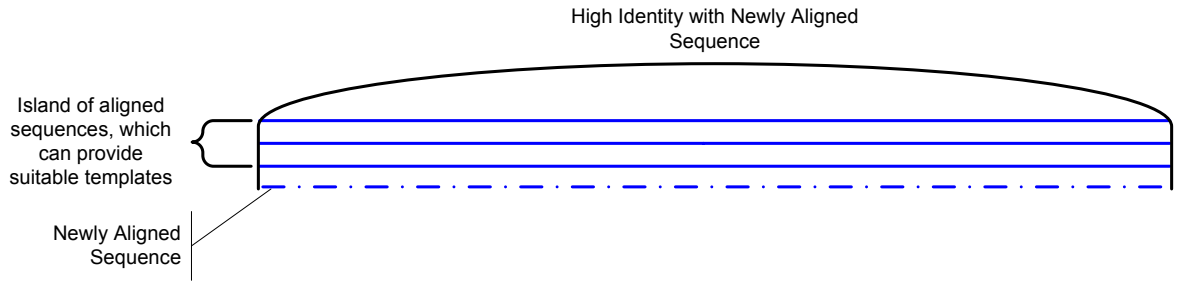


Figure 3.9: Schematic example of aligning a newly identified Category 1 RNA sequence within an existing "island"

Based on the definition of an "island", if the newly identified RNA sequence is a member of the identified "island" it will have high sequence identity with other members of the alignment that have already been aligned. Therefore, each member sequence that has already been aligned within the island represents a suitable "template" to guide the alignment of the newly identified RNA sequence.

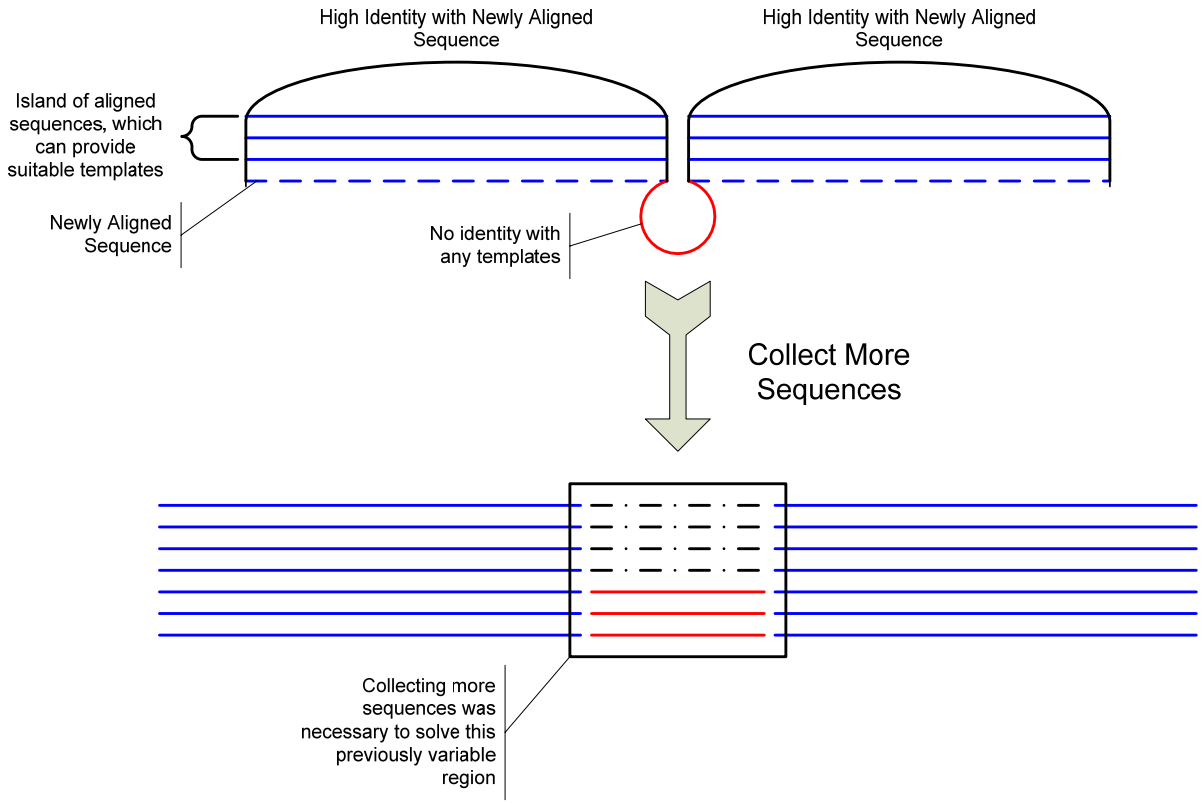


Figure 3.10: Aligning a Category 3 sequence with a region of "hypervariability"

Figure 3.3 introduced the concept the sequences within an "island" can exhibit regions of "hypervariability". The region of "hypervariability" has overlap with only a limited number of sequences within a given "island". Sequences which contain regions of "hypervariability" are considered Category 3 RNA sequences. This schematic is a pictorial representation of how a Category 3 RNA sequence is aligned. When the given RNA sequence is identified as a member of an "island" and contains a region of "hypervariability" is identified, existing sequences within the island are used as templates to align the sequence, excluding the region of "hypervariability". The region of "hypervariability" remains unaligned until more sequences are identified which overlap the existing sequence in the region of "hypervariability".

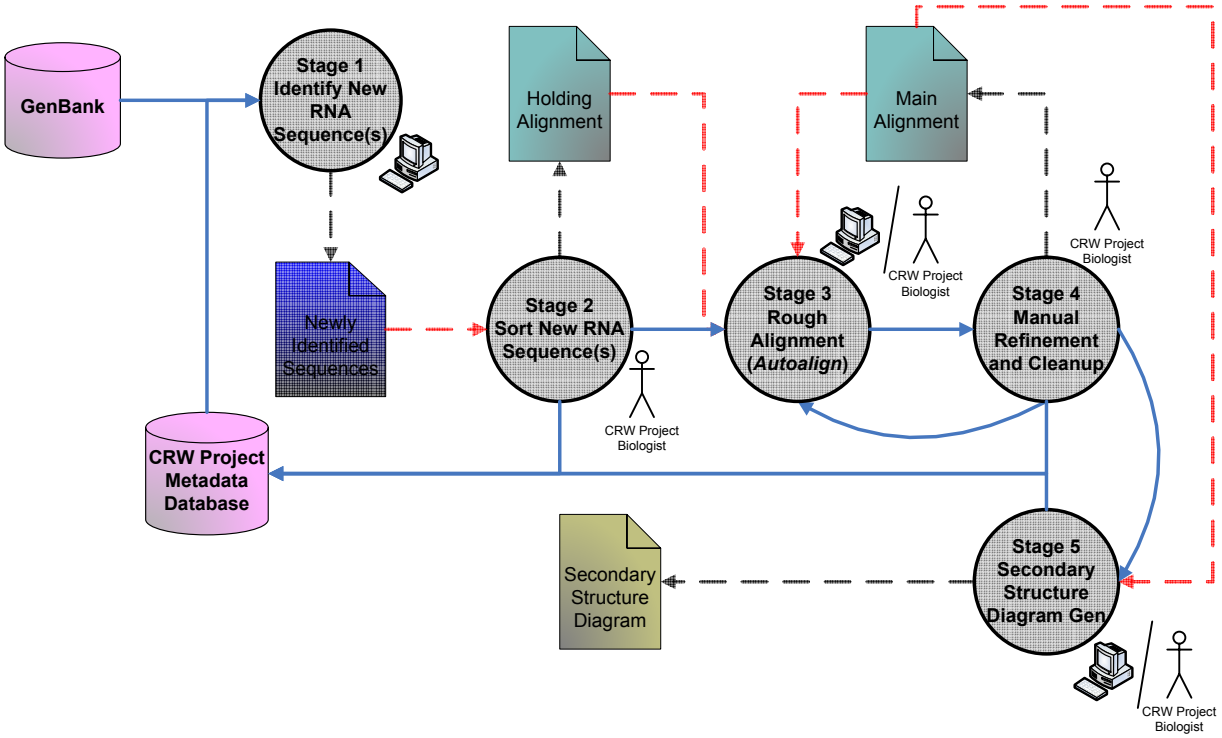


Figure 3.11: Schematic cartoon representation of the CRW Project *Curation Pipeline* circa 2002

The *Curation Pipeline* is the data management and curation workflow for the CRW Project. In this schematic, gray circles represent analysis steps, file icons represent data files which serve as input to or output from a given analysis step. Dark blue lines represent primary data flow pathways, dashed red lines indicate data flow from a file to an analysis step, dashed gray lines represent data flow from an analysis step to a file. Fully-automated analysis steps have a computer symbol adjacent to them. Fully-manual analysis steps have a stick figure adjacent to them. Semi-automated analysis steps have both a computer symbol and a stick figure adjacent to them. The *Curation Pipeline* is described in detail in Section 3.B.5.3.

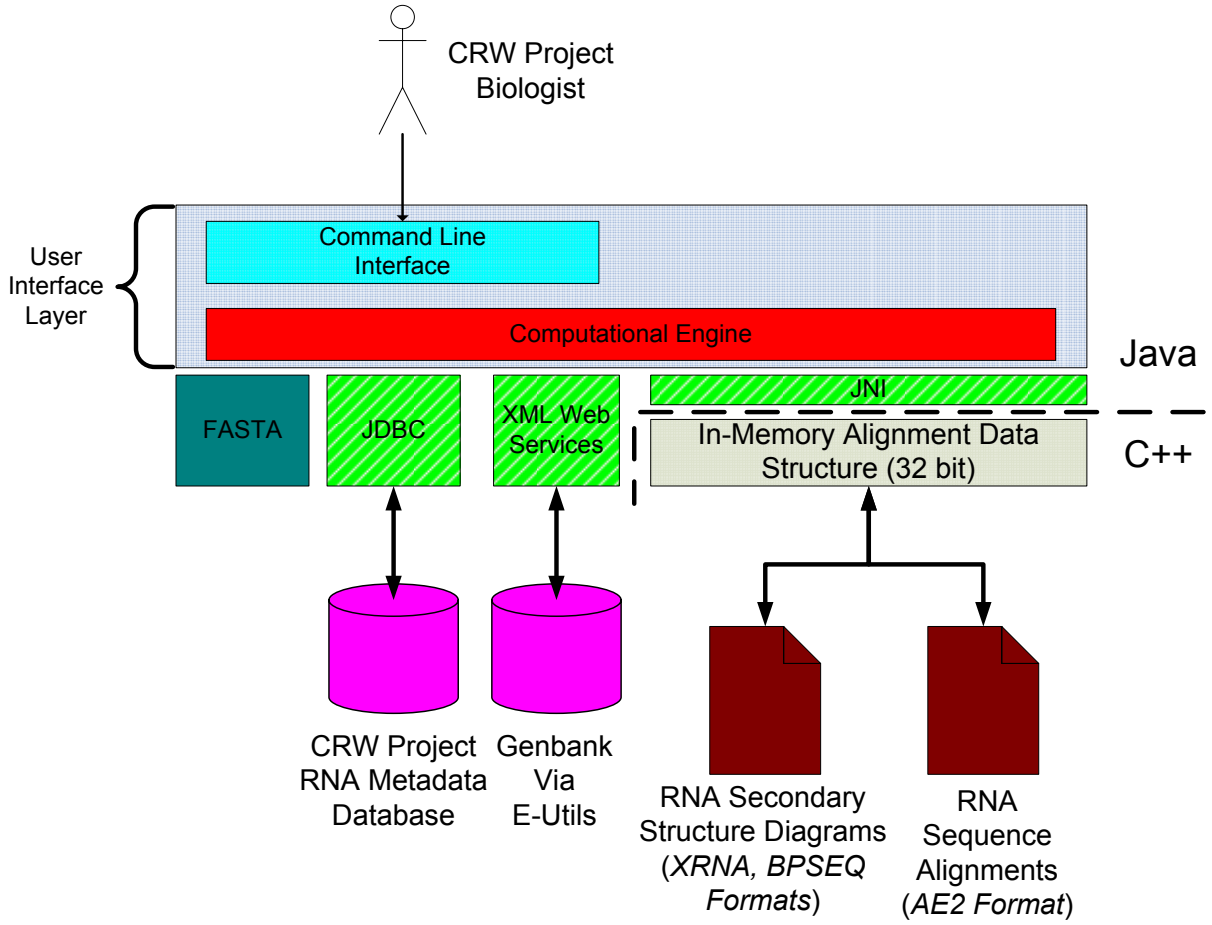


Figure 3.12: High level architecture diagram for the Comparative Analysis Toolkit (CAT)

CAT reads and writes RNA sequence alignments in AE2 (Section 3.B.2) format and RNA secondary structure diagrams in XRNA (Section 3.B.2) and BPSEQ format. BPSEQ format is an ASCII text format developed by the CRW Project for representing RNA secondary structure base pairings. CAT communicates directly with the CRW Project RNA Metadata Database (Section 3.B.5) through the Java Database Connectivity (JDBC) API and Genbank [129] through E-Utils. The CRW Project biologist interacts with CAT through a command-line interface. CAT is implemented in Java with the exception of the in-memory RNA sequence alignment data structures which are implemented in C++. The in-memory RNA sequence alignment data structures are linked with the remainder of the CAT application via the Java Native Interface (JNI). CAT can directly invoke FASTA [134-136] processes through internal Java APIs.

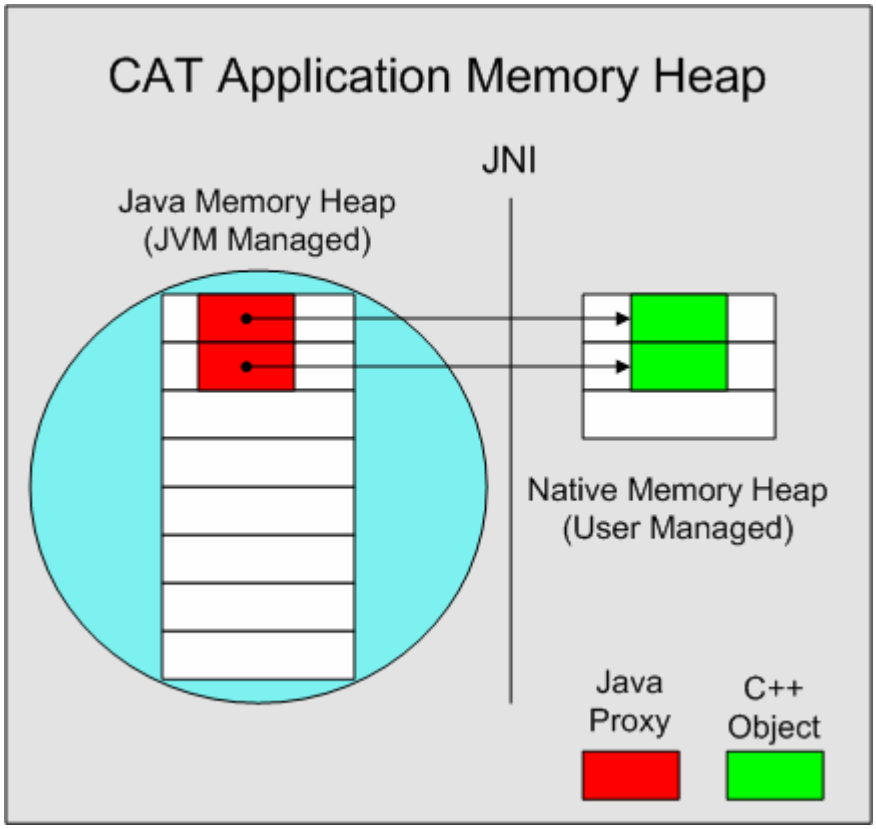


Figure 3.13: CAT application memory layout

CAT is a Java application which utilizes native objects implemented in C++ at runtime via the Java Native Interface (JNI). When any Java objects are created by CAT, they are created on the memory heap managed by the Java Virtual Machine (JVM). When CAT uses native objects, it creates a Java "proxy" which exists within the memory heap managed by the JVM. This Java "proxy" object holds a pointer to a native object which exists on a separate memory heap within the JVM, the "Native Memory Heap". The "Native Memory Heap" is managed by the user, not the JVM. As a result, native objects can be dynamically allocated and deallocated on the "Native Memory Heap" at anytime. In contrast, Java objects allocated on the memory heap managed by the JVM cannot be deallocated by the user on demand, rather the JVM contains collection "garbage collection" capabilities and automatically deallocates Java object on its own schedule. By using native objects, CAT has the ability to manage memory dynamically, which is necessary to support large in-memory RNA sequence alignments.

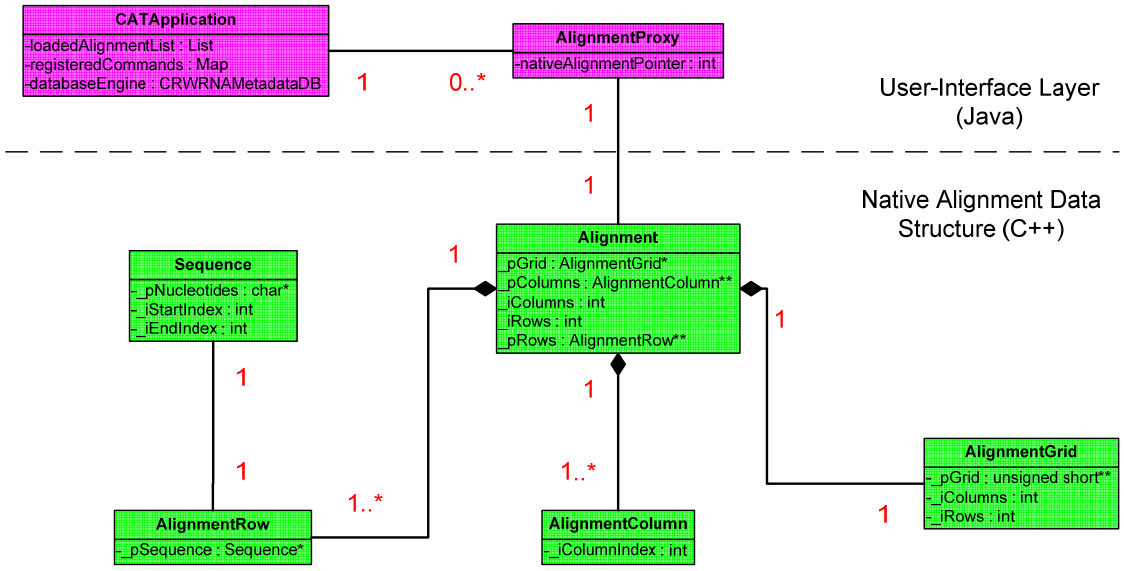


Figure 3.14: Unified Modeling Language (UML) object diagram which depicts the object model for the implementation of the C++ in-memory alignment data structure in CAT

Objects in green are implemented in C++ and objects in pink are implemented in Java. Red numbers on the object associations represent "multiplicities". For example the *AlignmentProxy* and *Alignment* objects are associated through a 1 to 1 relationship. The *CATApplication* and *AlignmentProxy* objects are represented by a 1 to many relationship; in this particular example 1 *CATApplication* object can associate with zero or more *AlignmentProxy* objects. In other words, the CAT application can hold more than one RNA sequence alignment in-memory at any given point in time. Associations with a diamond at the end represent "aggregations" which are whole-part relationships and define object lifetimes. For example, the lifetime of *AlignmentColumn* object is bounded by the Alignment object it is associated with. The *AlignmentColumn* object cannot exist without the corresponding *Alignment* object.

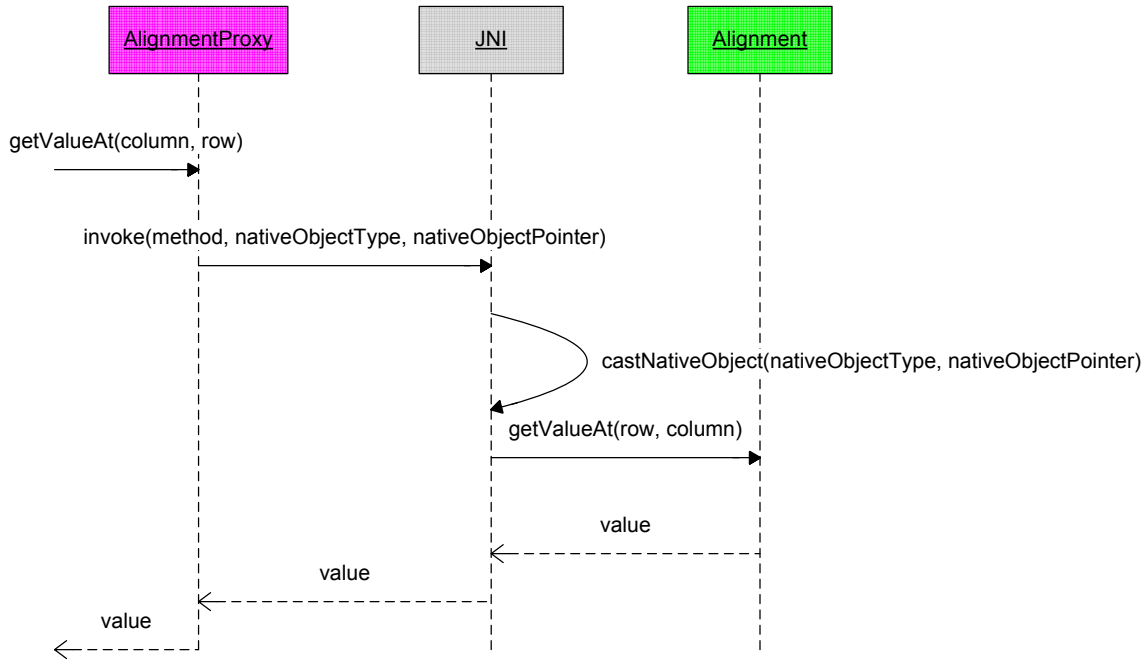


Figure 3.15: Unified Modeling Language (UML) sequence diagram which depicts how a given set of objects interacts in the implementation of a given use case

. The use case depicted in this sequence diagram is the retrieval of a value at a given column and row from the native in-memory sequence alignment. The call is made on the *AlignmentProxy* object. Using its reference to the native *Alignment* object, the *AlignmentProxy* object invokes the corresponding method on the *Alignment* object with the assistance of the Java Native Interface (JNI).

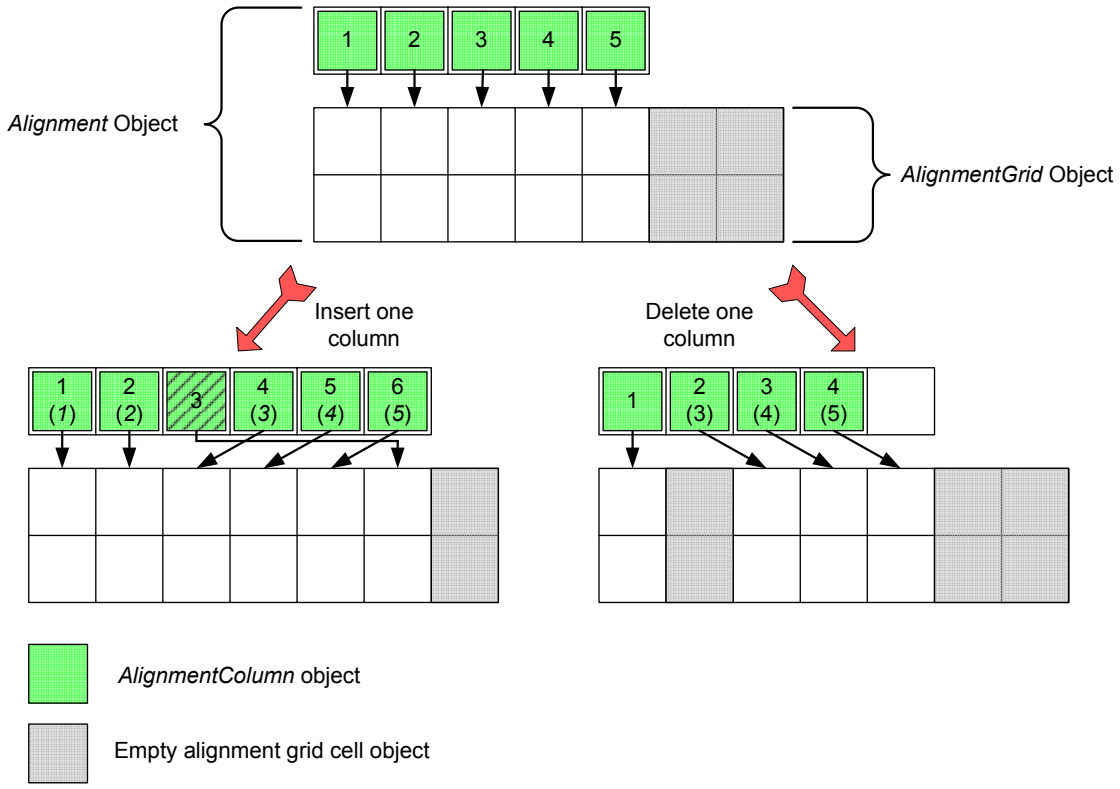


Figure 3.16: Schematic diagram of the novel indirection mechanism utilized by the native in-memory alignment data structure.

This indirection mechanism facilitates fast column insertion and deletion within a large in-memory RNA sequence alignment as well as direct access to any alignment grid cell object. The key to the indirection mechanism is that the order of the columns is not dictated by the *AlignmentGrid* object, which is implemented as a multi-dimensional array of unsigned shorts (Figure 3.14), rather the *Alignment* object has a separate one-dimensional array of *AlignmentColumn* objects. Each *AlignmentColumn* holds a pointer to a column in the *AlignmentGrid* that it corresponds to. All insertion and deletion of columns only occurs on the one-dimensional array of *AlignmentColumn* objects.

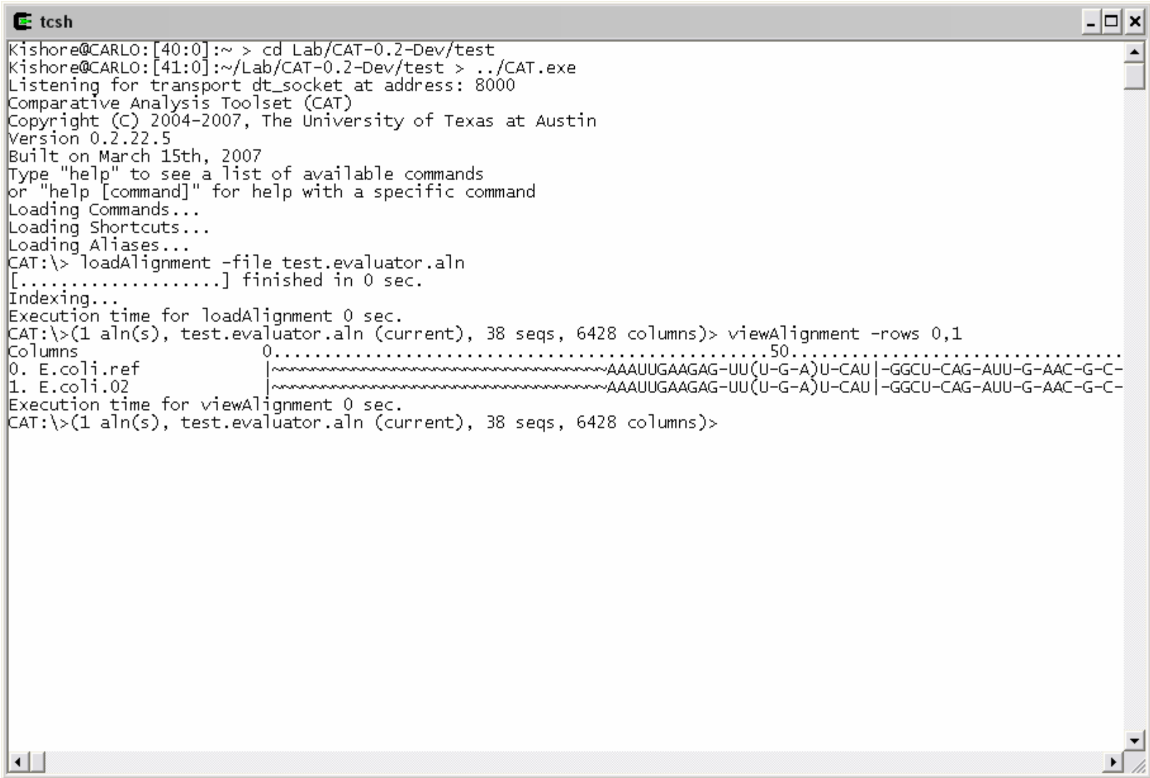


Figure 3.17: CAT application screen capture

One RNA sequence alignment is loaded in memory with the "*loadAlignment*" command, and the first two rows of that alignment are displayed with the "*viewAlignment*" command.

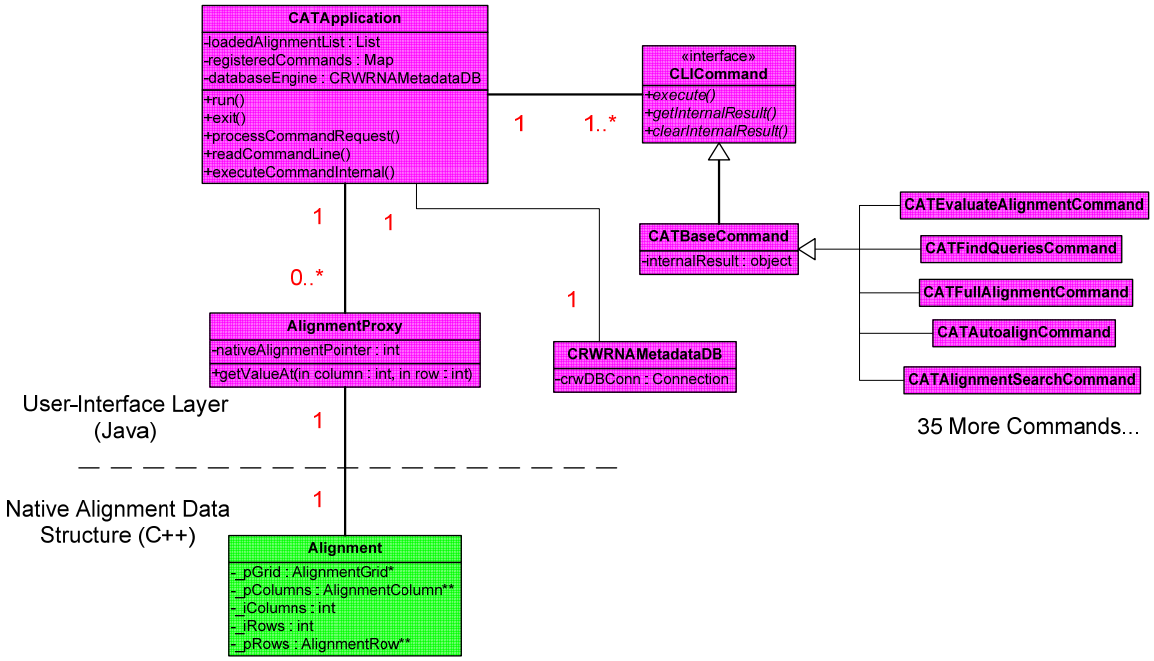


Figure 3.18: Unified Modeling Language (UML) object diagram which depicts the object model for the Java implementation of the core CAT application user interface layer.

Red numbers on the object associations represent “multiplicities” (Figure 3.14). Open ended arrows represent generalization associations. A generalization is equivalent to an “inherits” relationship. The main *CATApplication* object interacts with *AlignmentProxy* objects which represent in-memory RNA sequence alignments and the CRW RNA Metadata Database via the *CRWRNAMetadataDB* object. All CAT commands are derived from of the *CATBaseCommand* object which implements the *CLICommand* object interface. The CATBaseCommand object implements common functionality related to the CLICommand interface while the derived command objects such as the *CATAutoalignCommand* implement use case specific logic. The implication of using an interface to abstract all CAT commands is that the *CATApplication* object has no command specific functionality embedded within it.

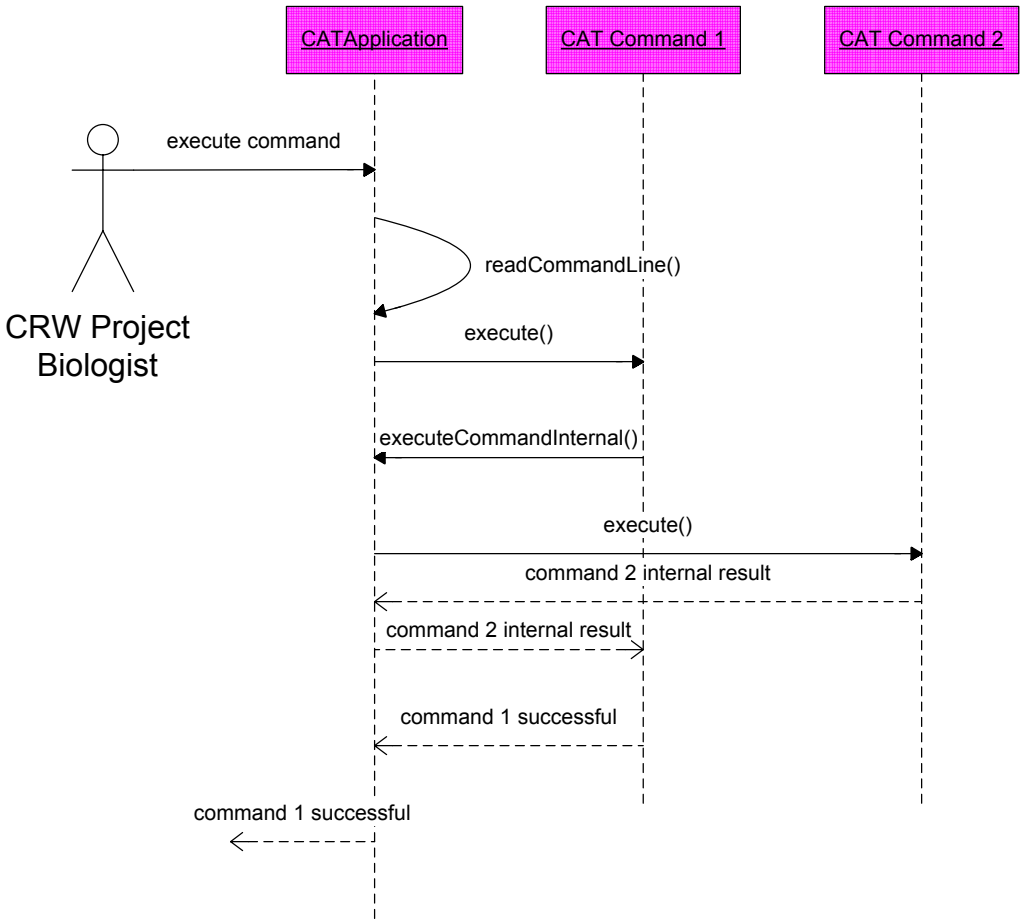


Figure 3.19: Unified Modeling Language (UML) sequence diagram which depicts the *Command Chaining* use case.

Command 1 and Command 2 can be any objects with implement the *CLICommand* interface (Figure 3.18). Using the **executeCommandInternal** function (*CATApplication*), any *CLICommand* can invoke another command within its processing flow via the *CATApplication* object. For example, the *CATAutoalignCommand* can invoke the *CATAlignmentSearch* Command and the results of the execution can be accessed by the *CATAutoalignCommand* through the **getInternalResult** function in *CLICommand* (Figure 3.18). A given *CLICommand* can execute an unlimited number of *CLICommand*(s) through *Command Chaining*, the only restriction is that a *CLICommand* may not call itself recursively.

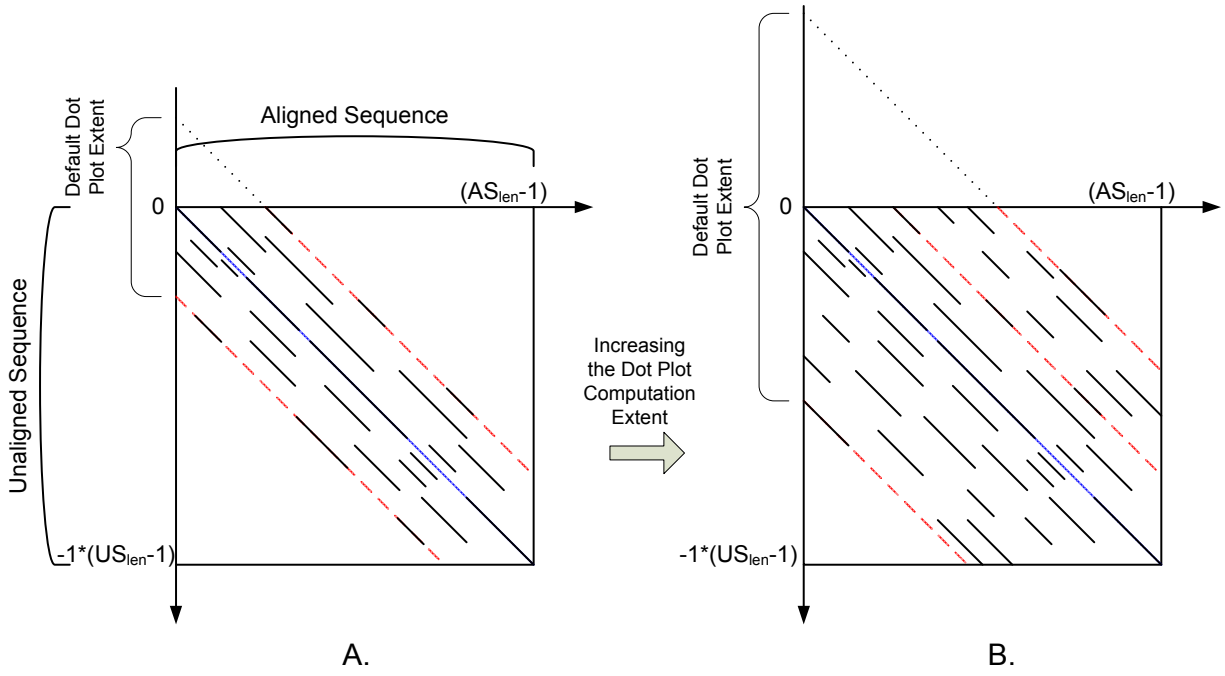


Figure 3.20: *Autoalign:* Computing a Partial Dot Plot

The first step in the "*autoalign*" heuristic pairwise alignment algorithm is to compute a partial dot plot to identify the longest line of similarity between the Unaligned Sequence (US) referred to as "query" and the Aligned Sequence (AS) referred to as "template". A line of similarity is any dark, diagonal line on the dot plot which indicates sequence identity between the "query" and the "template" sequence. For dot plots computed by "*autoalign*", the "template" sequence is plotted on the X axis from 0 to length-1 and the "query" is plotted on the Y axis from 0 to -1*(length-1). In dot plot **A.**, the dot plot is bounded by the two red, dashed diagonal lines, and the distance between their Y-intercepts along the Y axis is the dot plot extent. In dot plot **B.**, the dot plot extent is increased by widening the distance between the Y-intercepts of the two red, dashed diagonal lines. A dot plot extent of 100% requires the two red, dashed diagonal lines to have a Y-intercepts of $-1*(US_{len}-1)$ and $(AS_{len}-1)$ respectively.

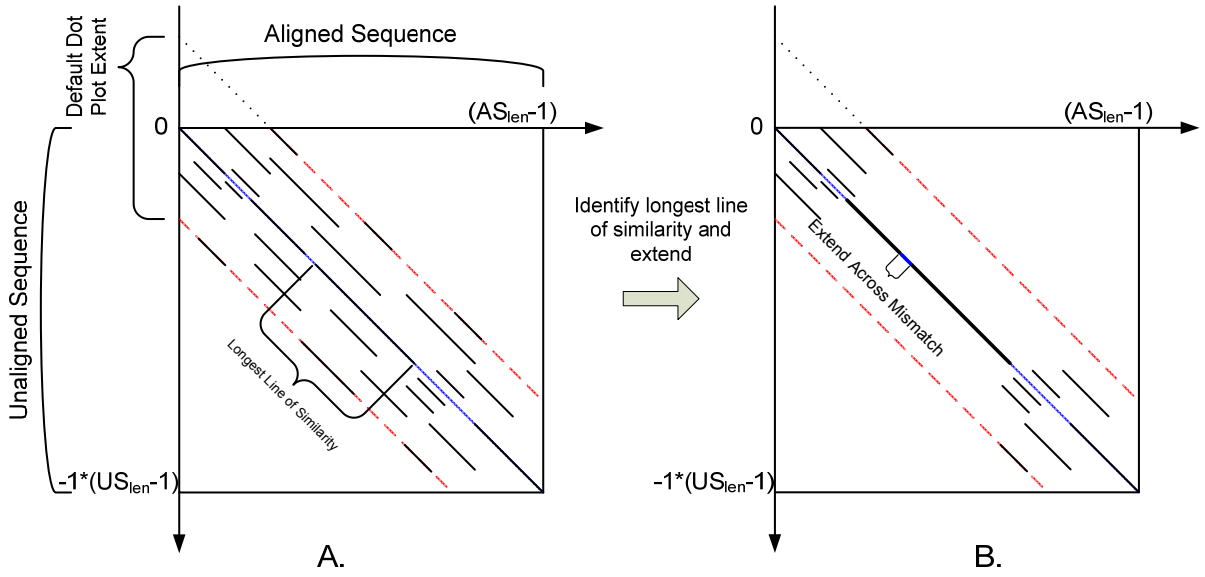


Figure 3.21: *Autoalign*: Extending the Longest Line of Similarity

The next step in the "autoalign" heuristic pairwise alignment algorithm is to extend the longest line of similarity in either direction. Similar to the dot plots in Figure 3.20, the Unaligned Sequence (US) is the "query" and is represented on the Y axis and the Aligned Sequence(AS) is the "template" and is represented on the X axis. The dot plot in **A.** marks the longest line of similarity. In dot plot **B.**, the longest line of similarity was extended by connecting it with a co-linear line of similarity that was longer than a given threshold and was separated from the longest line of similarity by a number of mismatches which was less than a given threshold.

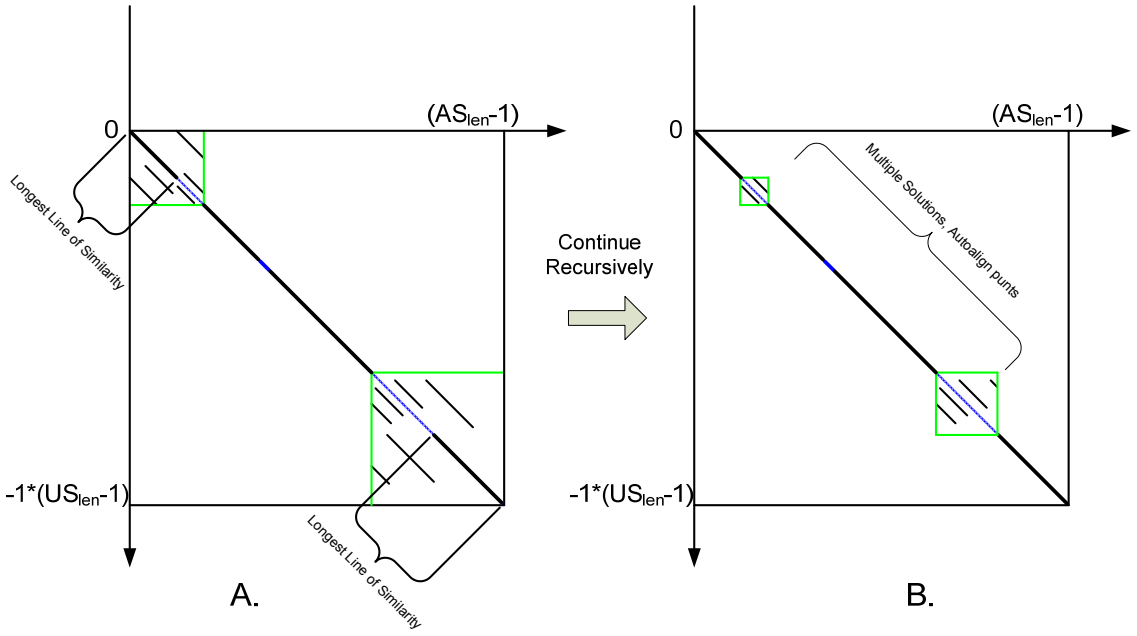


Figure 3.22: *Autoalign*: Recursive Step

The recursive step in the "*autoalign*" algorithm occurs when the longest line of similarity can no longer be extended. Similar to the dot plots in Figure 3.20, the Unaligned Sequence (US) is the "query" and is represented on the Y axis and the Aligned Sequence(AS) is the "template" and is represented on the X axis. In dot plot A. the longest line of similarity could no longer be extended (see Figure 3.20 and Figure 3.21 for the identification and extension of this line). Two new dot plots are computed (identified by the green boxes in **A.**) and the longest line of similarity is identified and extended as far as possible. The algorithm will continue computing smaller dot plots until the longest line of similarity can no longer be identified unambiguously on a given dot plot. It that situation, the algorithm stops as depicted on dot plot **B.** Two stop conditions are possible: 1) a given dot plot is empty (no similarity lines) or 2) a given has multiple similarity lines that are equivalent in length and longer than all other similarity lines on the dot plot.

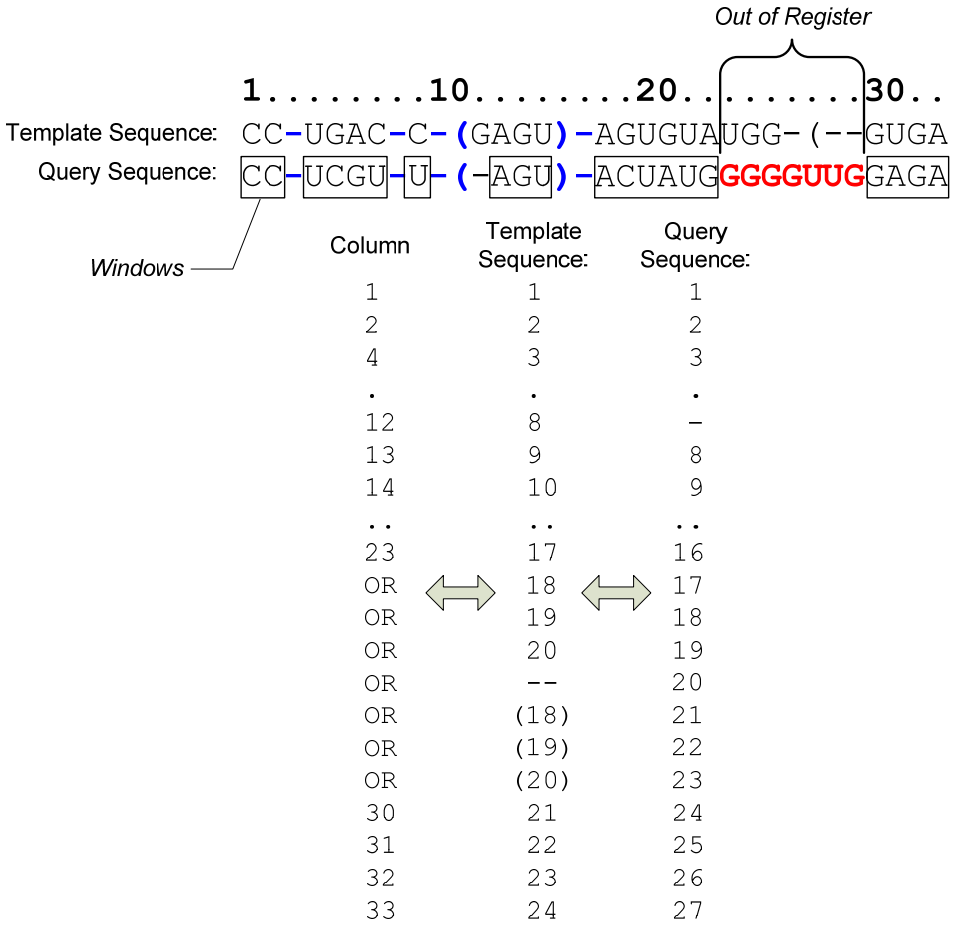


Figure 3.23: The second phase of the *autoalign* algorithm involves translating the newly aligned sequence into the existing RNA sequence alignment.

A column mapping is computed between the "template" and "query" based on the results of the pairwise alignment. (see Figures 3.20-3.22). The goal of the mapping is place nucleotides from the "query" in the same physical column of the alignment as the "template" nucleotide they are aligned with. A sliding window approach is used to place the nucleotides from the "query" in the appropriate columns starting from column 0. Annotation is copied from the "template" row to the "query" row where possible (marked in blue), and the translation algorithm is not allowed to add columns into the middle of the alignment. If two "query" nucleotides are aligned, but the number of intervening columns in the alignment is smaller than the number of intervening nucleotides in the "query", then the translation algorithm goes "Out of Register". In this example nucleotide 21 in the "query" should be aligned to nucleotide 18 in the "template"; however, the four nucleotides in the "query" (17-20) cannot be accommodated to support that mapping without adding columns. As a result, the translation is Out of Register by 4.

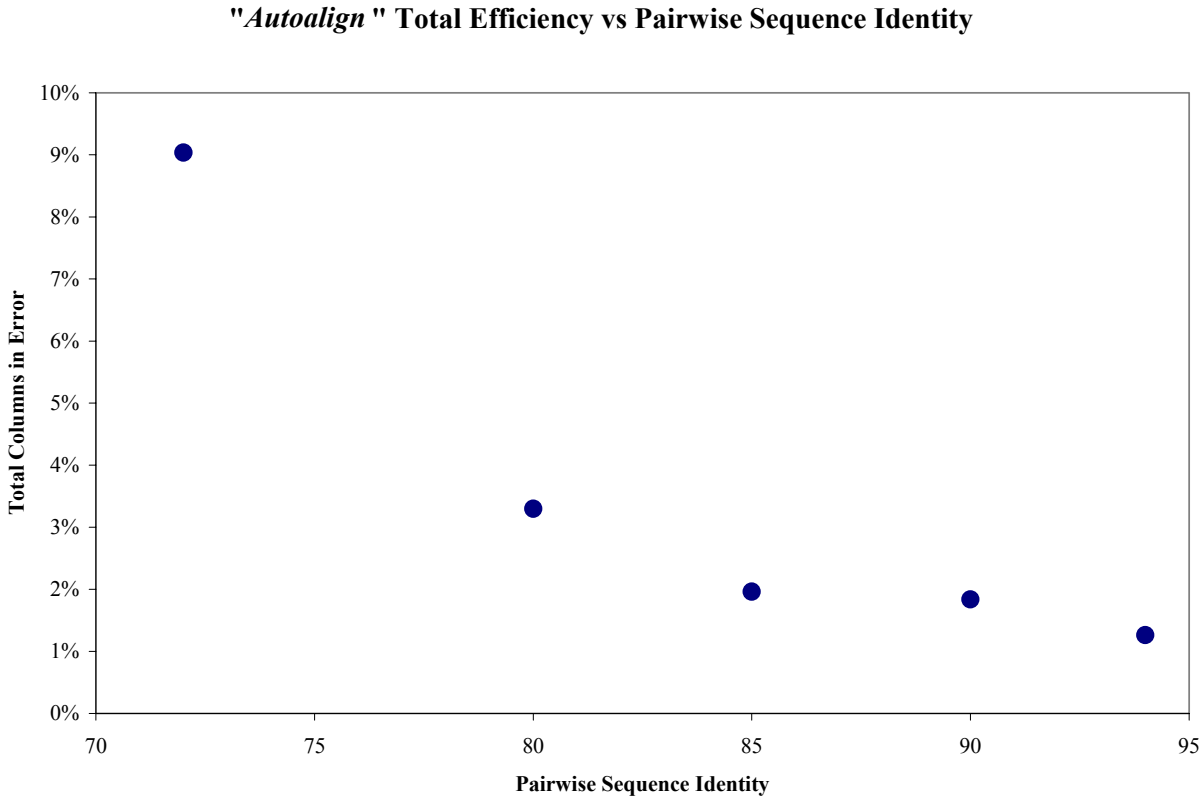


Figure 3.24: Characterizing the *autoalign* algorithm including the translation of the pairwise alignment result into an existing alignment.

To include the translation step (Section 3.C.2.1) in the characterization of the accuracy of the "*autoalign*" algorithm I consider the total number of errors on a per column basis for the "*autoalign*" result once incorporated into the existing alignment. The alignment results for five different "template" and "query" combinations are considered. The y-axis represents the percentage of total columns in the alignment that are in error after the alignment result is incorporated and the x-axis is the sequence identity between the "template" and the "query".

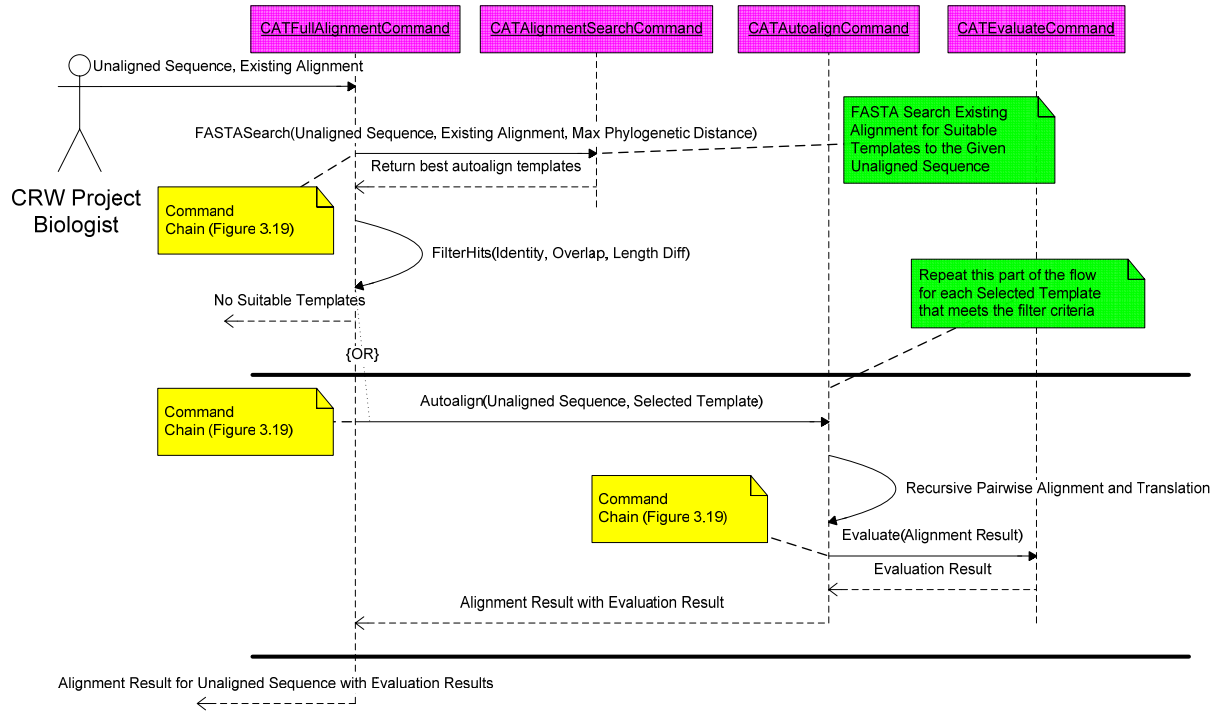


Figure 3.25: Unified Modeling Language (UML) sequence diagram which depicts the "*Full Alignment*" semi-automated RNA sequence alignment strategy as implemented in CAT

"Full Alignment" is a rigorous strategy to find the best possible template sequence to *autoalign* a given sequence. The CRW Project biologist first identifies an unaligned RNA sequence and an existing RNA sequence alignment. When the *CATFullAlignmentCommand* (Figure3.18) is invoked, CAT uses FASTA to search the existing RNA sequence alignment (Figure 3.11) to identify suitable template sequences to *autoalign* the unaligned RNA sequence. Using Command Chaining (Figure 3.19), the *CATAutoalignCommand* is invoked for each template sequence identified in the FASTA search. With each invocation, the *CATAutoalignCommand* invokes the *CATEvaluateCommand* (Section 3.C.3) via Command Chaining to provide the CRW Project biologist with a quantitative assessment of the alignment computed by *autoalign* with each candidate template sequence. The CRW Project Biologist picks the best alignment result.

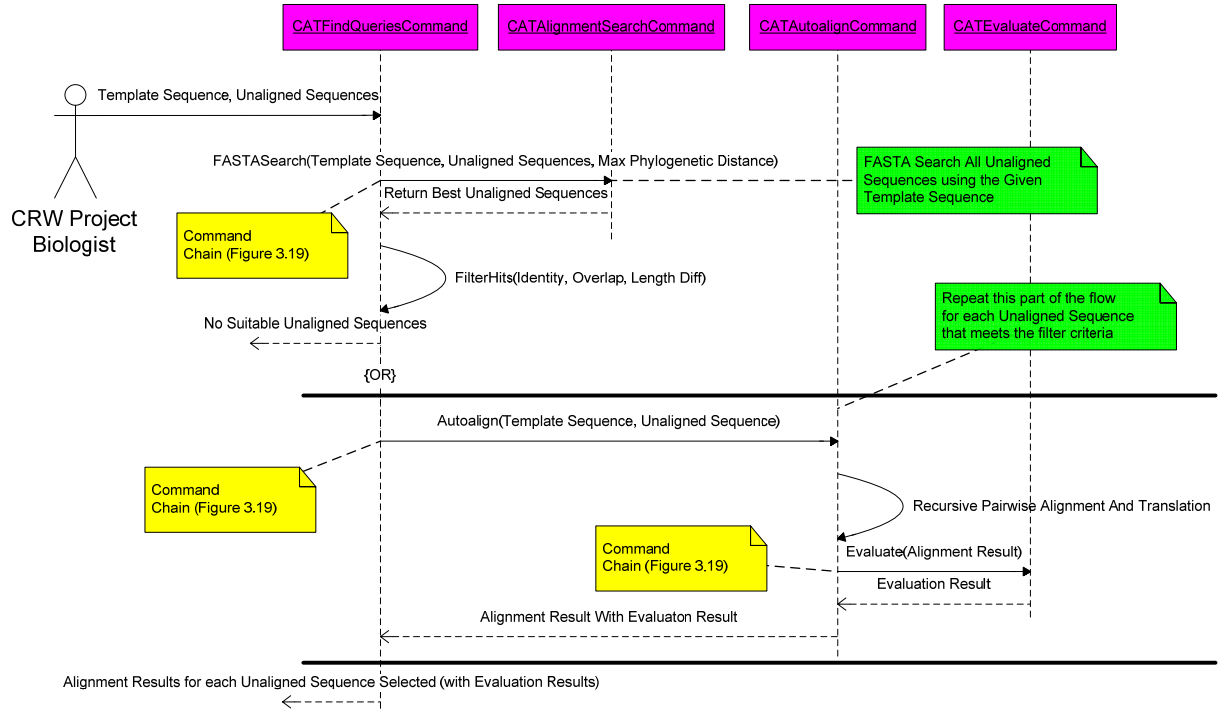


Figure 3.26: Unified Modeling Language (UML) sequence diagram which depicts the "Find Queries" semi-automated RNA sequence alignment strategy implemented in CAT.

In contrast to "*Full Alignment*", "*Find Queries*" is designed to aligning as many RNA sequences as possible with a given template sequence. The CRW Project biologist selects a sequence from the existing RNA sequence alignment as a template and a set of unaligned sequences from the appropriate holding alignment (Figure 3.12). Using a FASTA search, all unaligned sequences that meet the minimum criteria to be aligned using the template sequence and *autoalign* are selected. Similar to "*Full Alignment*" (Figure 3.26) the *CATAutoalignCommand* is invoked through Command Chaining to align each unaligned sequence identified. Contrary to "*Full Alignment*", the CRW Project Biologist only obtains one possible alignment result for each unaligned sequence; however, all possible unaligned sequences which could be aligned with the given template are aligned in one command invocation. For a given template sequence, hundreds of unaligned sequences can be aligned at once.

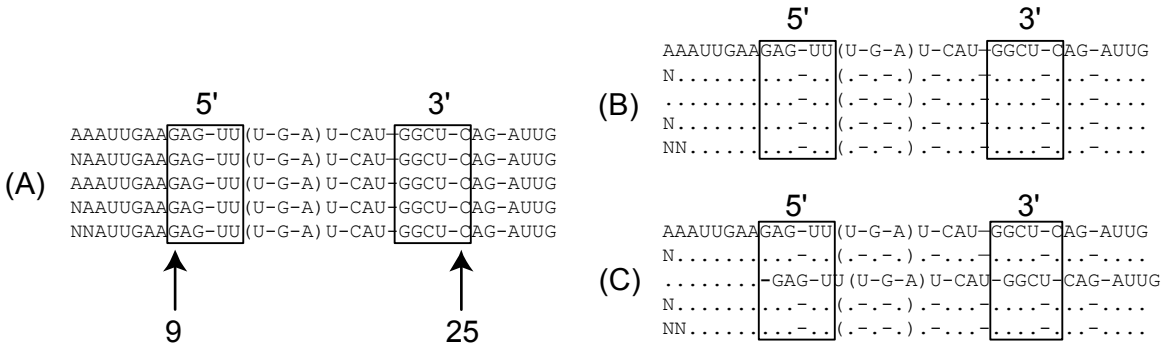


Figure 3.27: A highly conserved segment of the Bacterial 16S Ribosomal RNA (rRNA) alignment.

This well-characterized region has few degrees of freedom for juxtaposition of nucleotides. Panel **A** presents the sequence alignment with the 5' and 3' halves of a conserved secondary structure helix identified. Panel **B** is a "diff" view where any sequence which has the same nucleotide as first sequence (in a given column) is replaced with a '.', revealing the sequence conservation within the segment. In Panel **C**, an additional gap character is introduced in row 3 and a significant number of anomalies are immediately detected in the "diff" view.

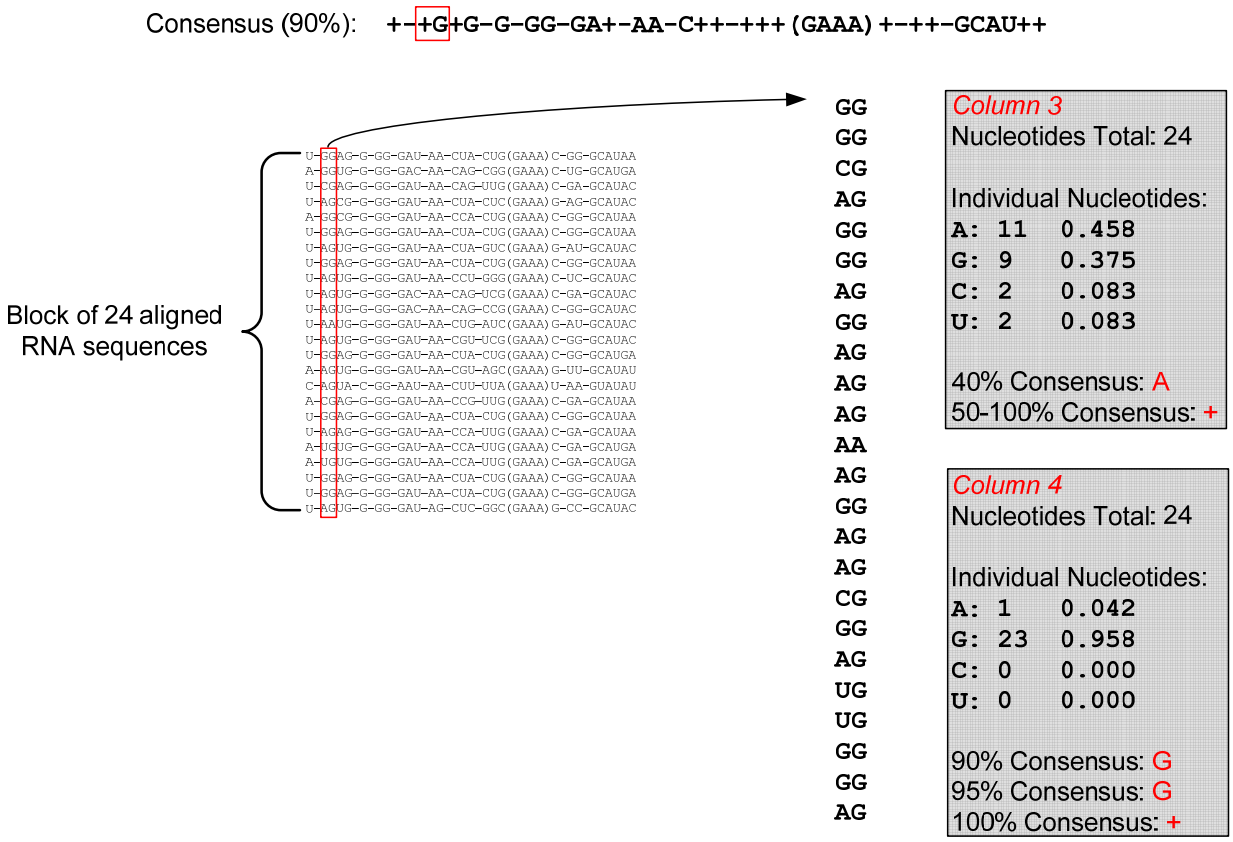


Figure 3.28: An example consensus sequence for a block of 24 aligned RNA sequences

Each column of the consensus can be represented with either: 1) an IUPAC nucleotide, 2) a '+', or 3) and annotation character. The '+' indicates that a specific nucleotide does not occur in a certain percentage of rows over which the consensus is computed. In this example, the 90% consensus is displayed. If an individual nucleotide is represented for a column in the 90% consensus then at least 90% of the rows over which the consensus was computed have that nucleotide in that column. If a '+' is used for a column in the 90% consensus, then at least 90% of the rows have a nucleotide, but not the **same** nucleotide. Nucleotide frequencies are calculated for each column in the specified block to determine a consensus sequence and an example computation for two columns within the block of 24 aligned RNA sequences is provided.

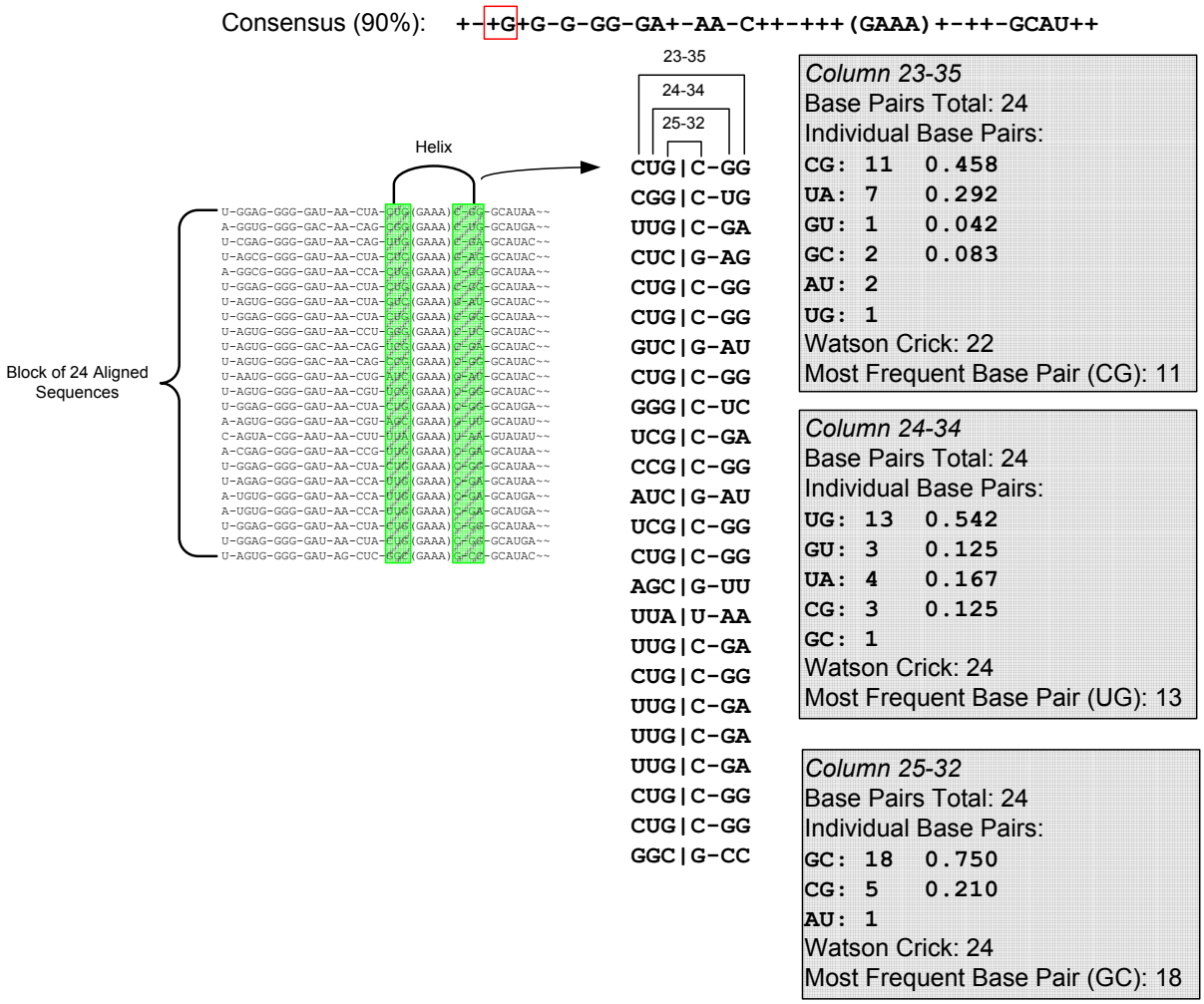
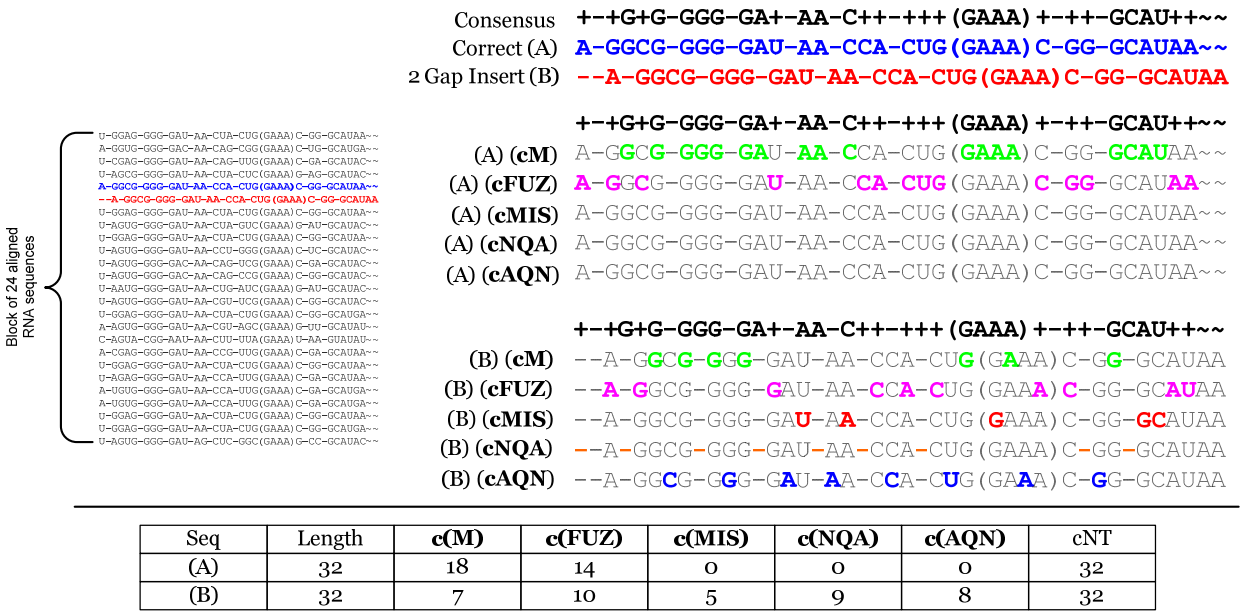


Figure 3.29: An example base pair frequency computation for a set of 3 base pairs projected across a block of 24 aligned RNA sequences

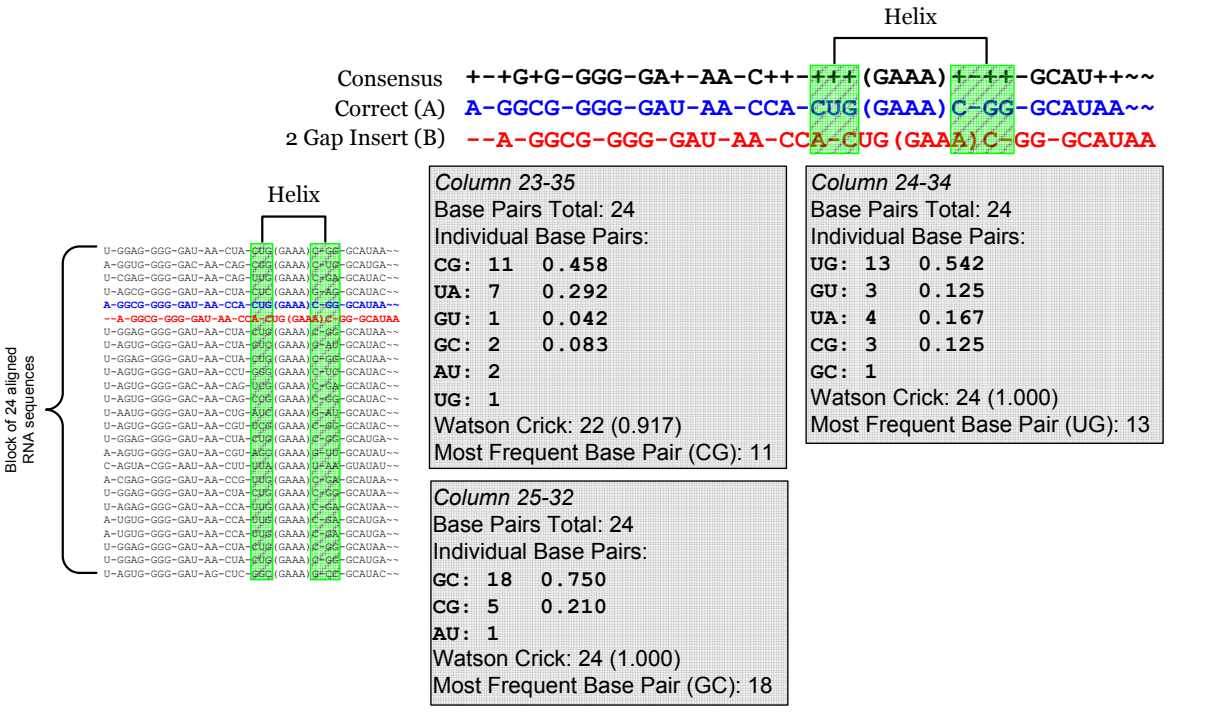
The same block of 24 aligned RNA sequences from Figure 3.28 is used. A three base pair secondary structure helix is highlighted in green and our goal is to characterize the sequence composition of those base pairs. The seven columns of the alignment within the helix are blown up to the right. Base pair frequencies are determined by looking down the two columns involved in the base pairing simultaneously. For base pair 23:35, 24 base pairs are observed, 11 of these base pairs are CG (46%). The total number of Watson crick + wobble base pairs observed for base pair 23:35 is 22.



Seq	Length	c(M)	c(FUZ)	c(MIS)	c(NQA)	c(AQN)	cNT
(A)	32	18	14	0	0	0	32
(B)	32	7	10	5	9	8	32

Figure 3.30: An example of sequence-based quality assessment for the alignment of a given RNA sequence using the "*evaluator*".

A sequence (**A**) is selected from a block of 24 aligned RNA sequences (Figure 3.28 and Figure 3.29) and two additional gap characters are inserted at the 5' end to create sequence (**B**). Nucleotides and annotation characters in sequence (**A**) and (**B**) are differentially colored depending on which of the five categories (**cM**, **cFUZ**, **cMIS**, **cNQA** or **cAQN**) (Section 3.C.3.1) they are associated with when compared against the consensus sequence. For example, Sequence (**A**) has 18 nucleotides that match the consensus exactly and are colored green. These matches are tabulated in the **cM** category. Sequence (**B**), which has a two gap insertion error has a significant number of mismatches to the consensus (**cMIS**) and a significant number of gaps and annotation characters which are aligned to nucleotides (**cAQN** and **cNQA**).



Seq	e(BPM)	e(BPWC)	s(BPD)	s(BPM)	s(BPWC)
(A)	2	3	3	2	3
(B)	1	1	0	0	0

Constants:
sBPM Default Threshold: 50%
sBPWC Default Threshold: 75%

Figure 3.31: An example of structure-based quality assessment for the alignment of a given RNA sequence using the "*evaluator*".

The same block of 24 aligned RNA sequences from Figure 3.30 is selected. Sequence (**A**) is the same sequence selected in Figure 3.30 and Sequence (**B**) is Sequence (**A**) with a 2 gap insertion at the 5' end. The three base pair helix first introduced in Figure 3.29 is indicated by the green shaded boxes, and the base pair frequencies are represented in the gray shaded boxes. The categories **s(BPD)**, **s(BPM)** and **s(BPWC)** are defined in Section 3.C.3.2. If we consider the projection of the helix over sequence (**A**) and (**B**), sequence (**A**) has the potential to form all three base pairs; however, sequence (**B**) can only form base pair 25:32. The most frequently occurring base pair is observed above 50% (**sBPM** default threshold) only for base pairs 24:34 and 25:32. Therefore the expected base pair match (**eBPM**) is two for sequence (A) and 1 for sequence (B). The expected Watson-crick base pair match count (**eBPWC**) is computed in a similar fashion.

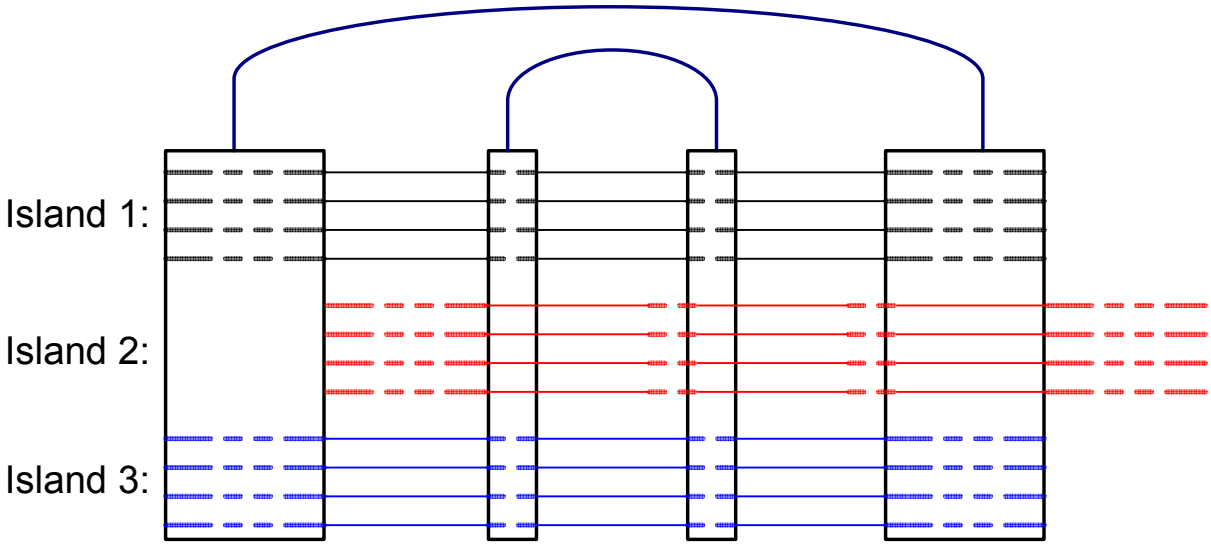


Figure 3.32: Abstract representation of an RNA sequence alignment as an example of how structure-based evaluation can detect misalignment.

Sequence samples from three different "Islands" 1, 2, and 3 are represented. Our definition of an "island" is that sequences within an "island" exhibit high sequence and structural identity. In contrast when sequences from two different "islands" are compared, significantly lower sequence identity is observed which structure identity is observed through common patterns of variation (Section 3.B.5.1). The location of common structure in this hypothetical RNA sequence alignment is identified by black connected boxes. All sequences from Island 2 are shifted such that they exhibit none of the expected structural relationships. The absence of these expected structural relationships would be observed in the evaluation of the alignment of sequences in Island 2 using structure-based evaluation.

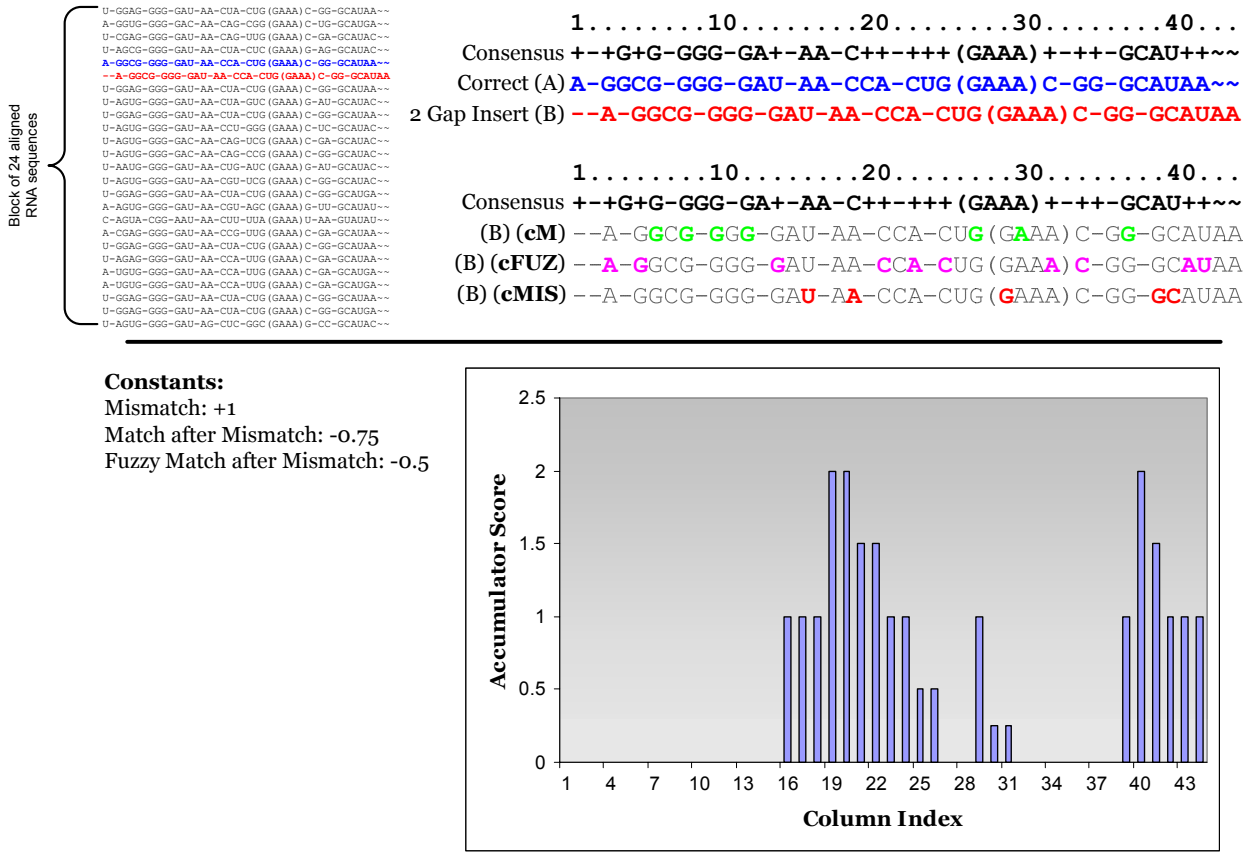
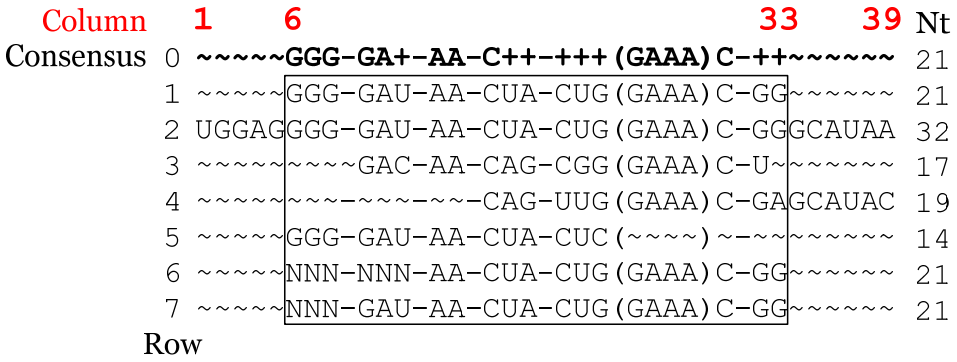


Figure 3.33: Detecting the locations of significant misalignment or "hotspots" from a sequence perspective by tracking error accumulation.

The same block of 24 aligned RNA sequences from Figure 3.30 is selected. Sequence (**A**) is the same sequence selected in Figure 3.30 and Sequence (**B**) is Sequence (**A**) with a 2 gap insertion at the 5' end. The "*MismatchAccumulationAnalyzer*" (Section 3.C.3.3) module in the "evaluator" tracks the accumulation of errors in sequence-based evaluation across the alignment. Each mismatch encountered increases the accumulator score by 1; each exact match following a mismatch decreases the accumulator score by 0.75; each fuzzy match following a mismatch decreased the accumulator score by 0.50. The plot above depicts the accumulator score (Y-axis) as a function of column index (X-axis) for Sequence (**B**).



Row	Real Length	Effective Length	Pct Complete
2	32	21	100
3	17	17	81
4	19	13	62
5	14	14	67
6	21	15	71
7	21	18	86

Figure 3.34: Detecting the locations of significant misalignment or "hotspots" from a sequence perspective by tracking error accumulation.

Computing the percent complete for a given sequence within the context of the RNA sequence alignment. In the hypothetical RNA sequence alignment above, the 90% consensus for an "island" (Section 3.B.5.1) of well-aligned, complete sequences (not shown) is represented in row 0. The consensus sequence establishes the expected 5' and 3' ends for any RNA sequence that is a member of the "island". The bounding box as determined by the consensus sequence is drawn around the seven sequences in this hypothetical RNA sequence alignment. The *Effective Length* is determined by counting the number of nucleotides for any sequence that are within the bounding box. For example, Sequence 2 has 21 nucleotides within the bounding box and 11 nucleotides outside the bounding box. The *Effective Length* of Sequence 2 is 21 nucleotides and its *Percent Complete* is 100% based on the consensus sequence. In the table, the *Effective Length* and *Percent Complete* for sequences 2 through 7 are reported.

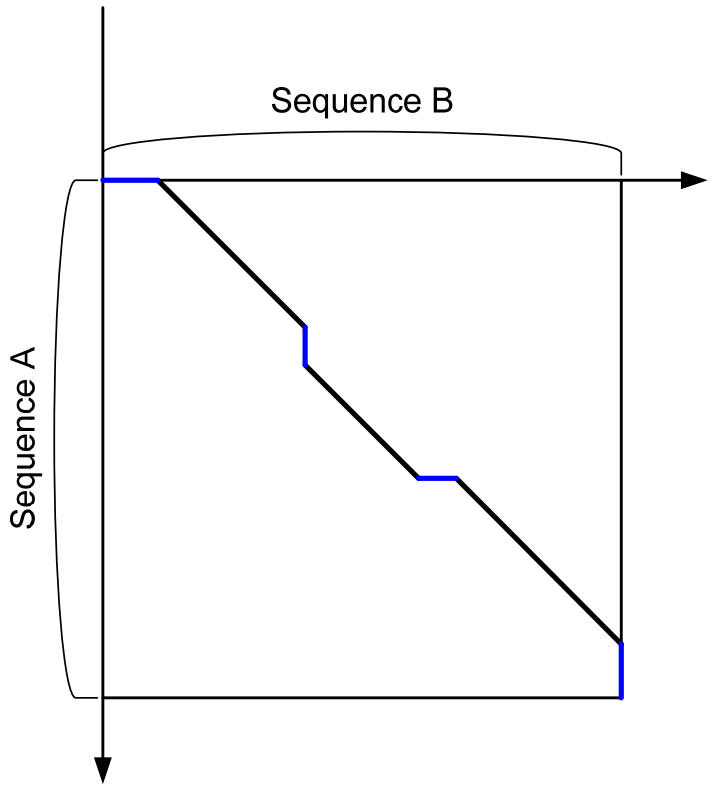


Figure 3.35: Hypothetical example of a maximal sequence alignment computed by Needleman and Wunsch or Smith and Waterman.

Blue lines represent gaps. A combination of acceptable substitution matrices based on observed evolutionary distance and penalties for opening and extending gaps are used to rigorously find the maximal alignment between two sequences with dynamic programming.

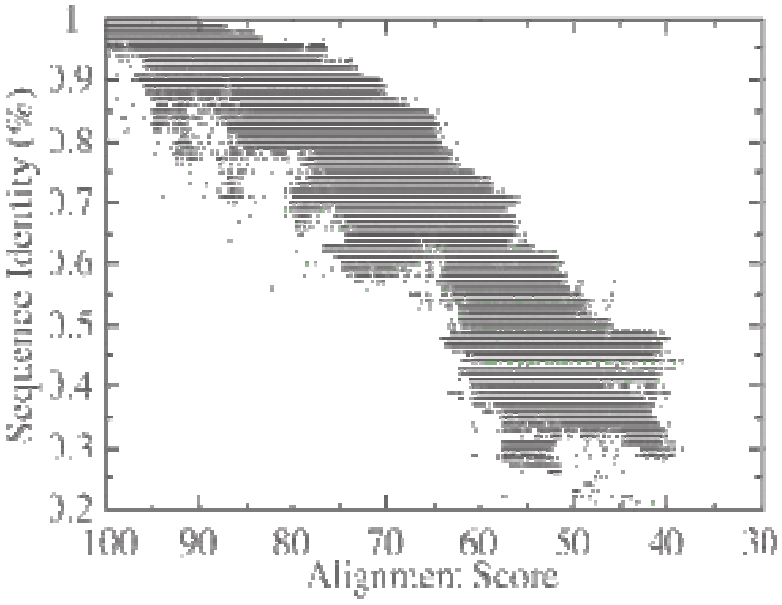


Figure 3.36: The Accuracy of a Clustal Generated RNA Sequence Alignment Compared to a Manually generated Alignment

The accuracy of a Clustal generated RNA sequence alignment is compared quantitatively against the RNA sequence alignment generated with the manual expert system approach introduced in Section 3.B.5 for a set of 800 small subunit animal mitochondrial small subunit Ribosomal RNA sequences (Eargle and Gutell, unpublished data). Clustal reproduces the manual based alignment with good accuracy for sequences that are identical with one another (i.e., within the same "island" (Section 3.B.5.1)); however, it does poorly when reproducing the alignment between sequences with significantly higher amounts of sequence variation (i.e., between two "islands" (Section 3.B.5.1)).

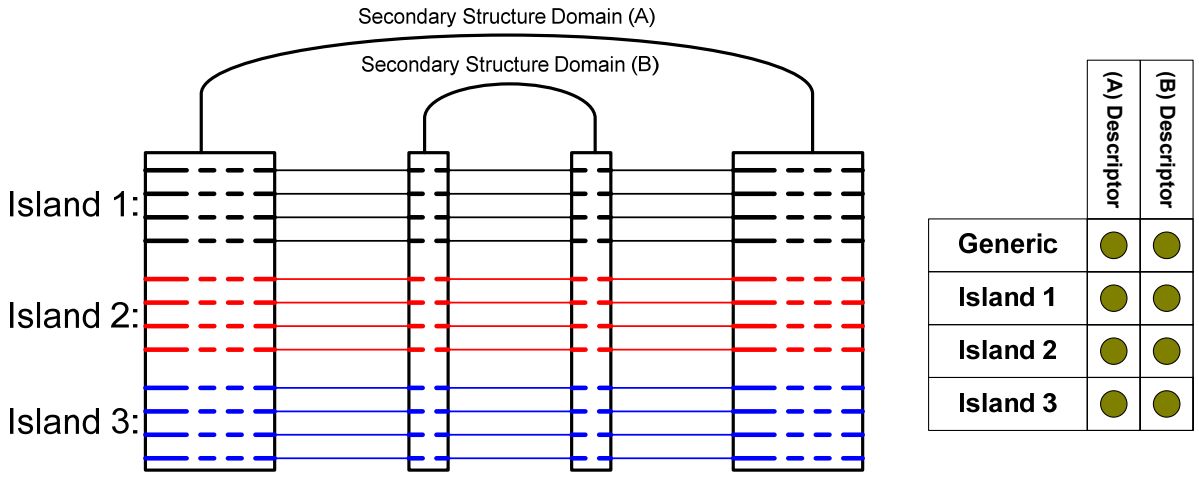


Figure 3.37: Abstract representation of an RNA sequence alignment to illustrate how a library of structural descriptions for RNAMOT or RNAMotif could be constructed.

Sequence samples from three different "Islands" 1, 2, and 3 are represented (Section 3.B.5.1). The location of common structure in this hypothetical RNA sequence alignment is identified by black connected boxes. Structural domains (A) and (B) apply to sequences in Islands 1, 2 and 3; therefore a generic descriptor which defines the base structural arrangement without a significant number of sequence constraints can be defined for domains (A) and (B). Within Islands 1, 2, and 3 the generic descriptors for domains (A) and (B) can be refined with sequence constraints. The total descriptor library for this hypothetical RNA sequence alignment would include 8 descriptors.

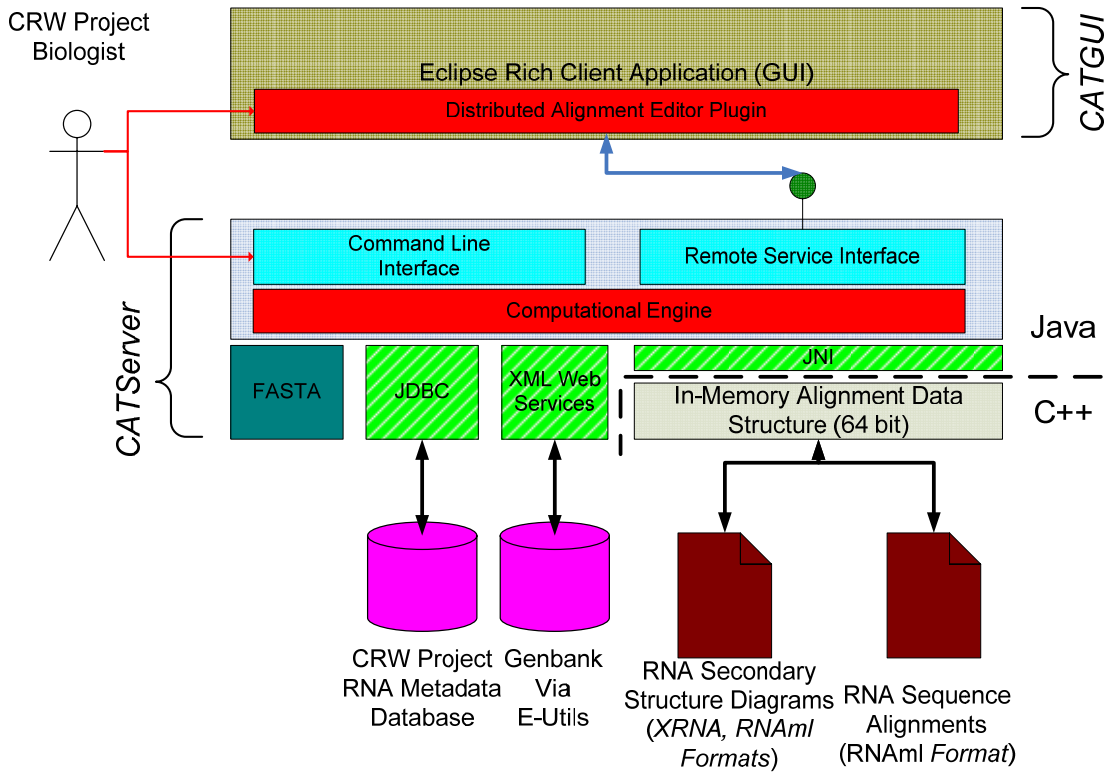


Figure 4.1: High level architecture diagram for the Comparative Analysis Toolkit (CAT) version 0.3.

This new architecture represents significant modifications and enhancements to the base CAT application architecture diagram presented in Figure3.12. The most important changes include: 1) the division of CAT into two components, *CATServer* and *CATGUI*; 2) the addition of a programmatic interface for remote software clients to interact with the *CATServer*; 3) distributed alignment visualization and editing with CATGUI, analysis results can be directly overlaid on RNA sequence alignment; 4) 64-bit support for the native in-memory alignment data structure in the CATServer, which can facilitate alignments of 10^6 or more sequences; 5) Secondary structure diagrams and RNA sequence alignments are primarily stored in RNAml (ref) format. Uses will still have the option to work with CAT 0.3 via the command-line interface in the *CATServer*. The *CATGUI* will be implemented as an Eclipse Rich Client application (http://www.eclipse.org) which will promote rapid development leveraging the GUI widgets and infrastructure already developed. Eclipse Rich Client applications can execute on Unix/Linux, Windows, and Mac OSX.

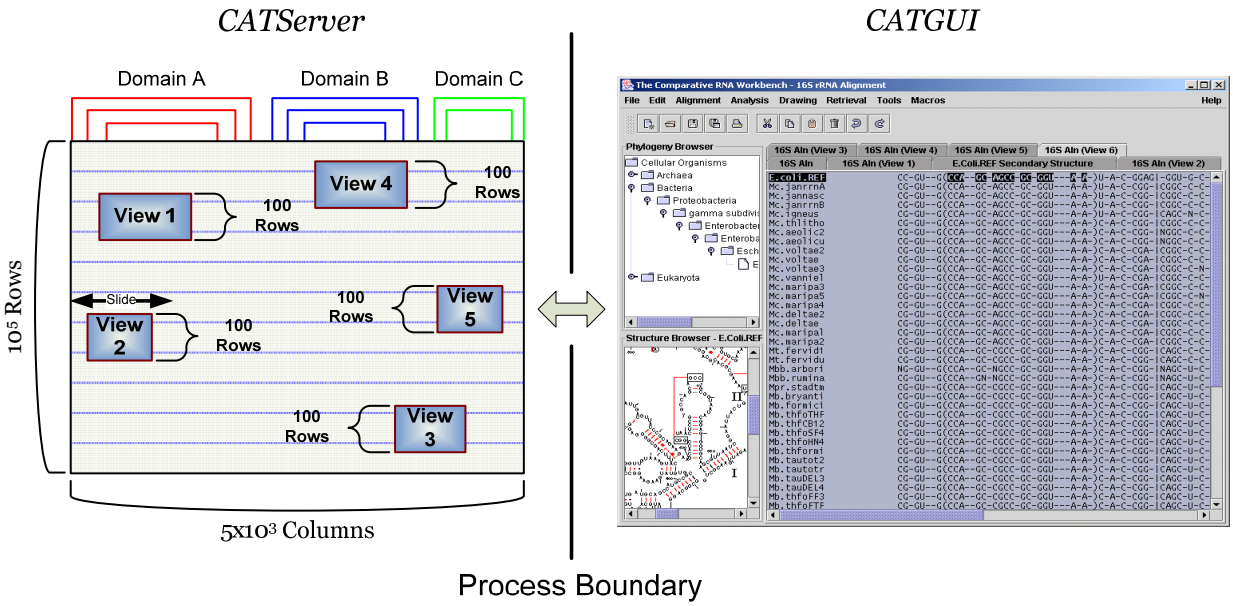


Figure 4.2: Schematic illustration of the distributed alignment editor concept for the *CATGUI*.

The entire RNA sequence alignment is loaded into memory in the *CATServer* process. The *CATGUI* provides a visualization of specific views of the alignment loaded in-memory in the *CATServer*. In this example, six separate views of 100 rows each are loaded into different tabbed windows within the *CATGUI*. As a result, the *CATGUI* has a significantly smaller memory footprint than the *CATServer*, which is crucial for good performance. Each of the views illustrated on the in-memory alignment in the *CATServer* are capable of sliding horizontally across the alignment. As the user scrolls a specific view, the view moves horizontally and the required data is streamed remotely from the *CATServer* to the *CATGUI*.

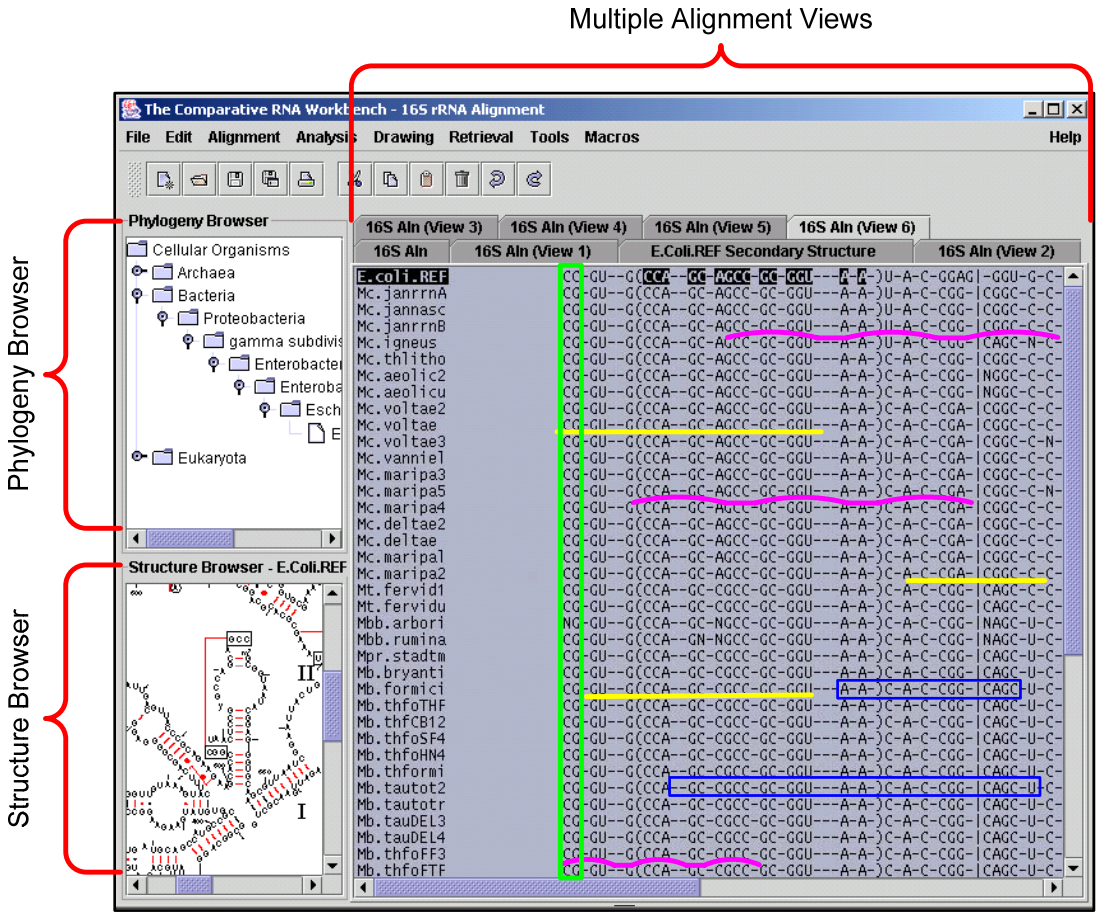


Figure 4.3: Close-up view of the proposed *CATGUI* implemented as a Java Swing Application.

Different alignment views (Figure 4.2) are located within tabbed windows for easy switching. Within any alignment view, the "evaluator" (Section 3.C.3) results can be annotated directly on the alignment. In this example, different potential annotation styles are represented. A "Structure Browser" is provided to facilitate navigation horizontally within a given alignment. Selecting the 5' or 3' nucleotide of a base pair on the "Structure Browser" will automatically scroll the alignment view to that column in the alignment. A "Phylogeny Browser" is provided to facilitate navigation vertically within a given alignment. Selecting a given node in the phylogenetic tree will result in a new alignment view centered on that node.

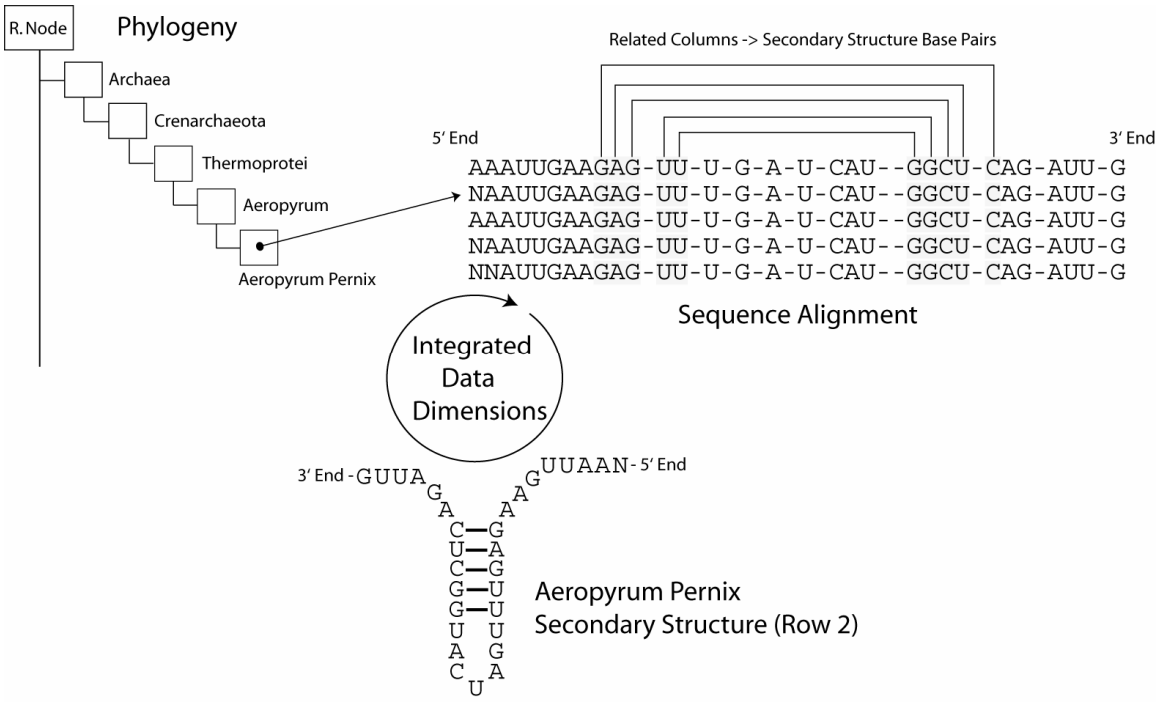


Figure 4.4: Schematic representation of the three primary data dimensions (phylogeny, secondary structure and sequence)

Schematic representation of the three primary data dimensions (phylogeny, secondary structure and sequence) to be considered in the design of a data model and database schema to facilitate RNA comparative analysis from an expert systems perspective. The database model must capture the relationships between rows (i.e., phylogeny) and columns (i.e., secondary structure) in the alignment and facilitate database queries that combine these dimensions.

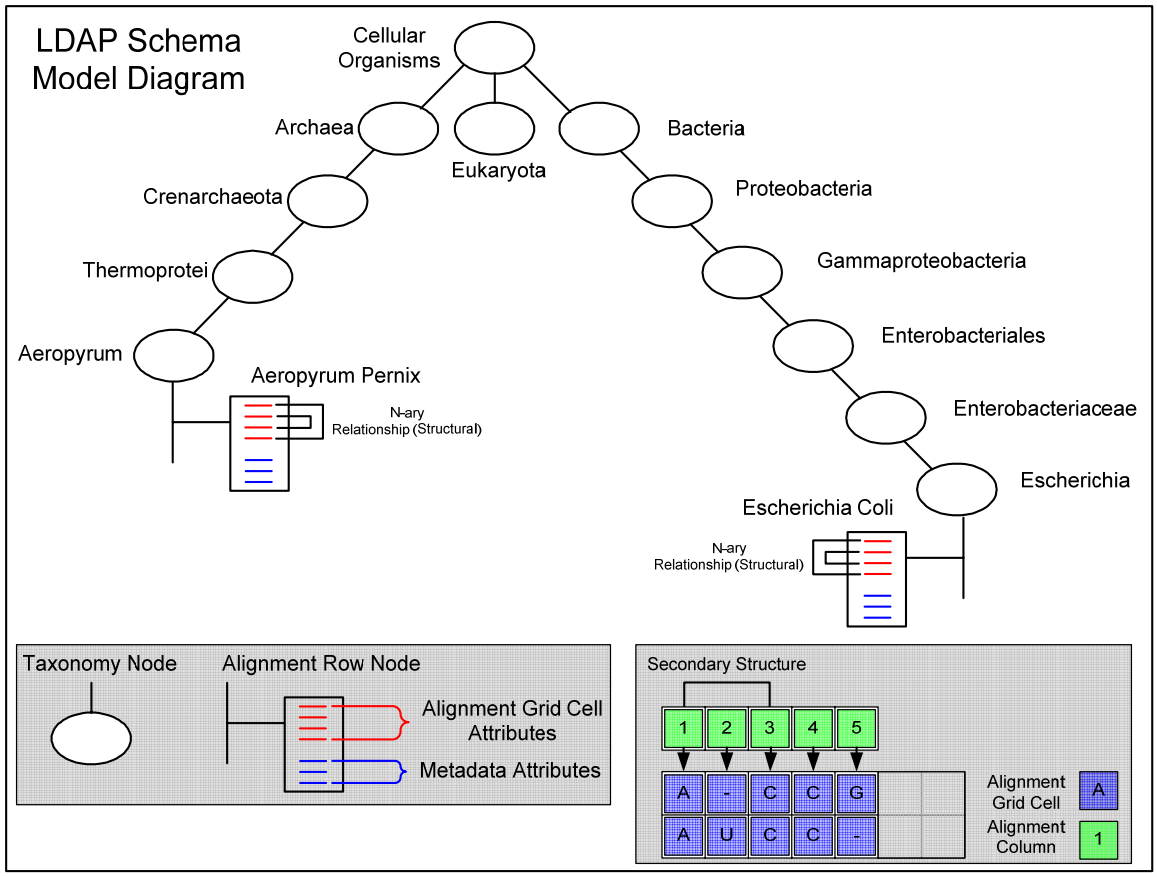


Figure 4.5: Schematic diagram of the data model used to persist an RNA sequence alignment with its associated phylogenetic and structural relationships in the Sun One Directory Server (SODS) hierarchical database

The basic schema followed the hierarchy of the phylogenetic tree published by the NCBI taxonomy project (ref). Each node of the phylogenetic tree is represented by a "Taxonomy Node" object in the database. Individual sequences in the alignment are represented as "Alignment Row Nodes". Each "Alignment Row Node" contains one attribute for each "Alignment Grid Cell" and the number of "Alignment Grid Cell" attributes within the "Alignment Row Node" is equivalent to the number of columns in the alignment. Furthermore, each "Alignment Row Node" contains extra metadata information about the sequence similar to what is contained in the current CRW Project RNA Metadata database (Section 3.B.5). Structural relationships are modeled as n-ary relationships between different "Alignment Grid Cell" attributes.

Appendix A

A.1 Statistics of the Comparatively Predicted Structure Database

Columns:

Accession: Genbank Identifier

Length: Total Nucleotides

AGCU: Count of all A, G, C and U Nucleotides

Other: Count of all other Nucleotides besides A, G, C and U

NPA: Not Publically Available (Accessions Column)

Web Reference:

http://www.rna.ccbb.utexas.edu/SIM/4C/mfold_Eval/seq_biases

Transfer RNA (tRNA)	Accessions	Length	AUGC	Other
Archaea				
Archaeoglobus fulgidus (Ala:A)	AE000965	75	72	3
Halobacterium salinarum (His:H)	X03198	76	75	1
Halobacterium salinarum (Asn:N)	X03195	76	76	0
Halobacterium salinarum (Gln:Q)	X03196	76	75	1
Halobacterium salinarum (Val:V)	K02505	78	78	0
Halobacterium salinarum (Val:V)	K00244	78	78	0
Halobacterium sp. NRC-1 (Ala:A)	AE005128	75	72	3
Halobacterium sp. NRC-1 (Cys:C)	AE005128	79	76	3
Halobacterium sp. NRC-1 (Gly:G)	AE005077	74	74	0
Halobacterium sp. NRC-1 (Arg:R)	AE004980	76	76	0
Haloferax volcanii (Ala:A)	K02507	75	75	0
Haloferax volcanii (Ala:A)	K02506	75	75	0
Haloferax volcanii (Ala:A)	K02508	75	75	0
Haloferax volcanii (Cys:C)	X02584	80	80	0
Haloferax volcanii (Asp:D)	K00170	76	76	0
Haloferax volcanii (Glu:E)	K00190	78	78	0
Haloferax volcanii (Glu:E)	K02510	78	77	1
Haloferax volcanii (Phe:F)	K02511	77	76	1
Haloferax volcanii (Gly:G)	K02515	74	74	0
Haloferax volcanii (Gly:G)	K02513	74	74	0
Haloferax volcanii (Gly:G)	K02514	74	73	1
Haloferax volcanii (His:H)	K02516	76	76	0
Haloferax volcanii (Ile:I)	K00219	77	77	0

Haloferax volcanii (Lys:K)	K02518	77	75	2
Haloferax volcanii (Pro:P)	K02521	74	74	0
Haloferax volcanii (Pro:P)	K02522	76	76	0
Haloferax volcanii (Gln:Q)	K00183	76	76	0
Haloferax volcanii (Arg:R)	K00154	76	76	0
Haloferax volcanii (Arg:R)	K02524	78	76	2
Haloferax volcanii (Thr:T)	K02526	76	75	1
Haloferax volcanii (Thr:T)	K02525	75	75	0
Haloferax volcanii (Val:V)	K02527	78	77	1
Haloferax volcanii (Val:V)	K00245	78	78	0
Haloferax volcanii (Trp:W)	K02528	77	74	3
Haloferax volcanii (Tyr:Y)	K00268	77	76	1
Methanobacterium formicicum (Asp:D)	AF443995	77	74	3
Methanocaldococcus jannaschii (Glu:E)	U67517	78	75	3
Methanocaldococcus jannaschii (Phe:F)	U67517	77	74	3
Methanocaldococcus jannaschii (His:H)	U67517	75	75	0
Methanocaldococcus jannaschii (Ile:I)	U67517	78	75	3
Methanocaldococcus jannaschii (Pro:P)	U67537	77	77	0
Methanocaldococcus jannaschii (Gln:Q)	U67528	76	73	3
Methanocaldococcus jannaschii (Arg:R)	U67492	78	75	3
Methanocaldococcus jannaschii (Thr:T)	U67528	77	74	3
Methanocaldococcus jannaschii (Val:V)	U67538	77	74	3
Methanococcus maripaludis (Lys:K)	AF108356	77	74	3
Methanococcus vannielii (Ala:A)	X00083	76	73	3
Methanococcus vannielii (Asp:D)	X00916	78	75	3
Methanococcus vannielii (Pro:P)	X00916	78	78	0
Methanococcus vannielii (Thr:T)	X00916	76	76	0
Methanococcus vannielii (Thr:T)	X00916	75	72	3
Methanococcus vannielii (Thr:T)	X00916	77	74	3
Methanococcus vannielii (Tyr:Y)	X00916	77	74	3
Methanosaeta concilii (Ala:A)	X51423	76	73	3
Methanospirillum hungatei (Ala:A)	M19342	76	73	3
Methanothermobacter thermautotrophicus (Ala:A)	AE000940	77	74	3
Methanothermobacter thermautotrophicus (Gly:G)	X06787	78	78	0
Methanothermobacter thermautotrophicus (Asn:N)	X06788	78	77	1
Methanothermus fervidus (Ala:A)	M32222	77	74	3
Methanothermus fervidus (Asp:D)	M26977	75	72	3
Methanothermus fervidus (Glu:E)	M26978	78	75	3
Methanothermus fervidus (Ile:I)	M26978	77	74	3
Methanothermus fervidus (Lys:K)	M26977	77	74	3
Methanothermus fervidus (Asn:N)	M26978	77	74	3
Methanothermus fervidus (Pro:P)	M26977	78	75	3
Methanothermus fervidus (Thr:T)	M26977	76	73	3
Pyrobaculum aerophilum (Ala:A)	AE009773	75	72	3
Pyrobaculum aerophilum (Ala:A)	AE009773	75	72	3
Sulfolobus solfataricus (Phe:F)	AE006696	77	74	3
Sulfolobus solfataricus (Val:V)	X06054	78	75	3

Thermococcus sp. MZ12 (Ala:A)	AY017180	77	77	0
Thermofilum pendens (Gly:G)	X14835	77	77	0
Thermofilum pendens (Met:M)	X14835	79	79	0
Thermofilum pendens (Met:M)	X14835	77	74	3
Thermofilum pendens (Met:M)	X14835	77	77	0
Thermoplasma acidophilum (Met:M)	K00302	76	75	1

Bacteria	**Accessions**	**Length**	**AUGC**	**Other**
Acholeplasma laidlawii (Trp:W)	X15508	75	75	0
Acidithiobacillus ferrooxidans (Ala:A)	X07395	76	76	0
Aeromonas hydrophila (His:H)	X12977	76	76	0
Aeromonas hydrophila (Pro:P)	X12977	77	77	0
Aeromonas hydrophila (Arg:R)	X12977	77	77	0
Agrobacterium tumefaciens str. C58 (Ala:A)	AE009341	76	76	0
Bacillus halodurans (Trp:W)	AP001510	74	74	0
Bacillus megaterium (Glu:E)	AF142677	75	72	3
Bacillus megaterium (Lys:K)	AF142677	76	76	0
Bacillus megaterium (Lys:K)	AF142677	76	75	1
Bacillus megaterium (Lys:K)	AF142677	76	75	1
Bacillus megaterium (Arg:R)	AF142677	77	77	0
Bacillus sporothermodurans (Ala:A)	AF071855	76	76	0
Bacillus subtilis (Ala:A)	K00141	76	76	0
Bacillus subtilis (Cys:C)	Z99108	74	74	0
Bacillus subtilis (Phe:F)	K00637	76	76	0
Bacillus subtilis (Gly:G)	K00637	74	74	0
Bacillus subtilis (Gly:G)	K00637	74	74	0
Bacillus subtilis (Gly:G)	Z99108	74	74	0
Bacillus subtilis (His:H)	Z99108	76	76	0
Bacillus subtilis (His:H)	K00637	76	76	0
Bacillus subtilis (Ile:I)	K00637	77	77	0
Bacillus subtilis (Ile:I)	Z99104	76	76	0
Bacillus subtilis (Ile:I)	Z99104	77	77	0
Bacillus subtilis (Ile:I)	K00637	77	77	0
Bacillus subtilis (Ile:I)	Z99104	77	77	0
Bacillus subtilis (Met:M)	K00637	77	77	0
Bacillus subtilis (Met:M)	K00637	77	76	1
Bacillus subtilis (Met:M)	K00297	76	76	0
Bacillus subtilis (Asn:N)	K00637	75	75	0
Bacillus subtilis (Pro:P)	K00637	77	77	0
Bacillus subtilis (Gln:Q)	Z99108	75	72	3
Bacillus subtilis (Arg:R)	K00156	76	76	0
Bacillus subtilis (Thr:T)	Z99104	76	73	3
Bacillus subtilis (Thr:T)	K00637	76	75	1
Bacillus subtilis (Val:V)	K00637	76	76	0
Campylobacter coli (Ala:A)	AF146727	76	76	0
Caulobacter crescentus (Ala:A)	L00194	75	75	0
Cyanophora paradoxa (Ala:A)	M19493	76	73	3
Cyanophora paradoxa (Glu:E)	U30821	75	72	3

Cyanophora paradoxa (Gly:G)	X51421	76	73	3
Cyanophora paradoxa (Ile:I)	M19493	77	77	0
Cyanophora paradoxa (Ile:I)	M19493	77	77	0
Cyanophora paradoxa (Ile:I)	M19493	77	77	0
Cyanophora paradoxa (Ile:I)	M19493	77	74	3
Cyanophora paradoxa (Ile:I)	M19493	77	74	3
Escherichia coli (Ala:A)	K00139	76	76	0
Escherichia coli (Asp:D)	AJ316554	77	77	0
Escherichia coli (Glu:E)	X05359	75	75	0
Escherichia coli (Glu:E)	K00188	76	76	0
Escherichia coli (Phe:F)	AF461394	76	76	0
Escherichia coli (Met:M)	K00296	77	77	0
Escherichia coli (Pro:P)	U00039	77	77	0
Escherichia coli (Arg:R)	K00152	76	76	0
Escherichia coli (Thr:T)	V00334	76	76	0
Escherichia coli K12 (Val:V)	AE000262	77	77	0
Geobacillus stearothermophilus (Phe:F)	K00332	76	76	0
Geobacillus stearothermophilus (Val:V)	K01065	76	76	0
Haemophilus influenzae Rd (Gly:G)	U32698	76	76	0
Lactobacillus delbrueckii (Asp:D)	X15246	77	73	4
Lactobacillus delbrueckii (Glu:E)	X15246	73	68	5
Lactobacillus delbrueckii (Asn:N)	X15245	76	72	4
Lactobacillus delbrueckii (Pro:P)	X15245	77	74	3
Lactobacillus delbrueckii (Arg:R)	X15246	77	74	3
Lactobacillus delbrueckii (Val:V)	X15246	76	73	3
Lactobacillus sakei (Gly:G)	AF401668	75	72	3
Mycobacterium tuberculosis CDC1551 (Val:V)	AE007103	75	72	3
Mycoplasma capricolum (Cys:C)	X16746	75	75	0
Mycoplasma capricolum (Asp:D)	X16745	77	77	0
Mycoplasma capricolum (Glu:E)	X16748	76	76	0
Mycoplasma capricolum (Gly:G)	X16749	74	74	0
Mycoplasma capricolum (His:H)	X16750	76	76	0
Mycoplasma capricolum (Lys:K)	X16756	76	76	0
Mycoplasma capricolum (Met:M)	X16758	77	77	0
Mycoplasma capricolum (Asn:N)	X16744	76	76	0
Mycoplasma capricolum (Gln:Q)	X16747	75	75	0
Mycoplasma capricolum (Thr:T)	X16764	76	76	0
Mycoplasma capricolum (Thr:T)	X16765	76	76	0
Mycoplasma capricolum (Thr:T)	X16764	76	73	3
Mycoplasma capricolum (Val:V)	X16769	76	76	0
Mycoplasma capricolum (Trp:W)	X16766	75	75	0
Mycoplasma capricolum (Trp:W)	X16767	77	74	3
Mycoplasma capricolum (Trp:W)	X16767	76	76	0
Mycoplasma mycoides (Ala:A)	X03154	76	76	0
Mycoplasma mycoides (Gly:G)	M21590	74	74	0
Mycoplasma mycoides (Ile:I)	X03154	77	77	0
Mycoplasma mycoides (Ile:I)	Y00372	77	77	0
Mycoplasma mycoides (Met:M)	X03154	77	77	0

Mycoplasma mycoides (Pro:P)	X03154	77	77	0
Mycoplasma mycoides (Arg:R)	X03154	77	77	0
Mycoplasma pneumoniae (Gly:G)	AE000043	74	74	0
Mycoplasma pneumoniae (Lys:K)	AE000043	75	75	0
Mycoplasma pneumoniae (Gln:Q)	AE000043	75	75	0
Mycoplasma sp. (Phe:F)	X01305	76	76	0
Mycoplasma sp. PG50 (Lys:K)	X05660	76	76	0
Pectobacterium carotovorum subsp. (Ile:I)	AF448597	77	77	0
Photobacterium phosphoreum (His:H)	X12976	76	76	0
Pseudomonas aeruginosa (Gly:G)	AE004843	74	74	0
Pseudomonas aeruginosa (Thr:T)	AF331071	76	76	0
Pseudomonas aeruginosa (Thr:T)	AE004843	76	76	0
Pylaiella littoralis (Ala:A)	X14875	76	73	3
Pylaiella littoralis (Ile:I)	X14875	77	74	3
Rhodospirillum rubrum (Phe:F)	K00331	76	76	0
Salmonella typhimurium (Pro:P)	AE008893	77	77	0
Salmonella typhimurium LT2 (Ala:A)	AE008786	76	76	0
Salmonella typhimurium LT2 (Cys:C)	AE008895	74	74	0
Salmonella typhimurium LT2 (Glu:E)	AE008839	76	76	0
Salmonella typhimurium LT2 (Gly:G)	AE008904	74	74	0
Salmonella typhimurium LT2 (Gly:G)	AE008904	76	76	0
Salmonella typhimurium LT2 (Gly:G)	AE008883	76	76	0
Salmonella typhimurium LT2 (Gly:G)	AE008809	75	74	1
Salmonella typhimurium LT2 (His:H)	AE008789	76	76	0
Salmonella typhimurium LT2 (Lys:K)	AE008799	76	76	0
Salmonella typhimurium LT2 (Asn:N)	AE008883	76	76	0
Salmonella typhimurium LT2 (Pro:P)	AE008727	77	77	0
Salmonella typhimurium LT2 (Pro:P)	AE008727	77	77	0
Salmonella typhimurium LT2 (Gln:Q)	AE008829	75	75	0
Salmonella typhimurium LT2 (Gln:Q)	AE008883	75	74	1
Salmonella typhimurium LT2 (Gln:Q)	AE008710	75	75	0
Salmonella typhimurium LT2 (Arg:R)	AE008893	77	77	0
Salmonella typhimurium LT2 (Arg:R)	AE008893	77	77	0
Salmonella typhimurium LT2 (Thr:T)	AE008893	76	76	0
Salmonella typhimurium LT2 (Thr:T)	AE008762	76	76	0
Salmonella typhimurium LT2 (Thr:T)	AE008809	76	75	1
Salmonella typhimurium LT2 (Thr:T)	AE008881	76	75	1
Salmonella typhimurium LT2 (Val:V)	AE008762	77	77	0
Salmonella typhimurium LT2 (Val:V)	AE008809	76	76	0
Salmonella typhimurium LT2 (Trp:W)	AE008881	76	76	0
Spiroplasma melliferum (Ala:A)	X03715	76	76	0
Spiroplasma melliferum (Cys:C)	X03715	76	76	0
Spiroplasma melliferum (Asp:D)	X03715	77	77	0
Spiroplasma melliferum (Phe:F)	X03715	76	76	0
Spiroplasma melliferum (Ile:I)	X03715	77	77	0
Spiroplasma melliferum (Met:M)	X03715	77	77	0
Spiroplasma melliferum (Pro:P)	X03715	77	77	0
Spiroplasma melliferum (Arg:R)	X03715	77	77	0

Staphylococcus epidermidis (Gly:G)	K00199	75	75	0
Staphylococcus epidermidis (Gly:G)	K00200	74	74	0
Staphylococcus epidermidis (Asn:N)	AF269878	75	72	3
Staphylococcus epidermidis (Asn:N)	AF269878	74	71	3
Streptococcus agalactiae (Ala:A)	AF291419	76	73	3
Streptomyces coelicolor (Lys:K)	AL596030	77	74	3
Streptomyces coelicolor A3(2) (Cys:C)	AL157953	74	74	0
Streptomyces coelicolor A3(2) (Gly:G)	AL157953	76	73	3
Streptomyces coelicolor A3(2) (Asn:N)	AL163003	76	73	3
Streptomyces coelicolor A3(2) (Asn:N)	AL163003	76	73	3
Streptomyces coelicolor A3(2) (Val:V)	AL157953	75	72	3
Synechococcus sp. PCC 7002 (Phe:F)	K02680	76	75	1
Synechocystis sp. (Glu:E)	M19535	76	75	1
Thermus thermophilus (Gly:G)	X51824	76	76	0
Thermus thermophilus (Ile:I)	M25628	77	77	0
Thermus thermophilus (Thr:T)	X51824	76	76	0
Thermus thermophilus (Thr:T)	X51824	76	76	0
Tolypothrix distorta (Ala:A)	AY007689	76	76	0
Vibrio cholerae (Asn:N)	AE004132	77	77	0
Vibrio cholerae (Pro:P)	AE004107	77	77	0

Eukaryotic Chloroplast	**Accessions**	**Length**	**AUGC**	**Other**
Chlamydomonas moewusii (Thr:T)	X51398	77	74	3
Chlamydomonas reinhardtii (Ala:A)	J01395	76	73	3
Chlamydomonas reinhardtii (Cys:C)	X54407	75	72	3
Chlamydomonas reinhardtii (Glu:E)	X54408	76	75	1
Chlamydomonas reinhardtii (Glu:E)	L26266	76	73	3
Chlamydomonas reinhardtii (Gly:G)	J01399	76	73	3
Chlamydomonas reinhardtii (Trp:W)	X62566	76	73	3
Chlorella ellipsoidea (Ala:A)	X05693	76	76	0
Chlorella ellipsoidea (Arg:R)	X15090	77	74	3
Chlorella pyrenoidosa (Ile:I)	X03848	77	74	3
Codium fragile (Gly:G)	M26736	75	75	0
Codium fragile (Met:M)	M26737	77	76	1
Codium fragile (Arg:R)	M26738	77	76	1
Cyanidium caldarium (Lys:K)	D17791	75	72	3
Euglena gracilis (Ala:A)	X70810	76	73	3
Euglena gracilis (Cys:C)	X70810	75	72	3
Euglena gracilis (Asp:D)	X70810	76	73	3
Euglena gracilis (Asp:D)	K00173	75	75	0
Euglena gracilis (Phe:F)	X70810	76	73	3
Euglena gracilis (Phe:F)	K00340	76	74	2
Euglena gracilis (Phe:F)	K00341	76	75	1
Euglena gracilis (Gly:G)	X70810	75	72	3
Euglena gracilis (Gly:G)	X70810	76	73	3
Euglena gracilis (His:H)	X70810	75	72	3
Euglena gracilis (Ile:I)	X70810	77	74	3
Euglena gracilis (Lys:K)	X70810	76	73	3

Euglena gracilis (Met:M)	X70810	76	73	3
Euglena gracilis (Asn:N)	X70810	75	72	3
Euglena gracilis (Pro:P)	X70810	77	74	3
Euglena gracilis (Gln:Q)	X70810	75	72	3
Euglena gracilis (Arg:R)	X70810	77	74	3
Euglena gracilis (Thr:T)	X70810	75	72	3
Euglena gracilis (Val:V)	X70810	76	73	3
Euglena gracilis (Trp:W)	X70810	76	73	3
Glycine max (Met:M)	X07377	76	73	3
Glycine max (Val:V)	X07675	75	72	3
Guillardia theta (Arg:R)	AF041468	76	73	3
Lactuca sativa (His:H)	AF426317	77	74	3
Marchantia polymorpha (Ala:A)	M20942	76	73	3
Marchantia polymorpha (Asp:D)	X04465	77	74	3
Marchantia polymorpha (Glu:E)	X04465	76	73	3
Marchantia polymorpha (Glu:E)	X04465	76	73	3
Marchantia polymorpha (Phe:F)	X04465	76	73	3
Marchantia polymorpha (Gly:G)	X01647	74	71	3
Marchantia polymorpha (Gly:G)	M20952	74	71	3
Marchantia polymorpha (His:H)	X04465	77	74	3
Marchantia polymorpha (Ile:I)	X04465	77	74	3
Marchantia polymorpha (Ile:I)	M20955	75	72	3
Marchantia polymorpha (Ile:I)	M20955	77	74	3
Marchantia polymorpha (Lys:k)	M20959	75	72	3
Marchantia polymorpha (Met:M)	X04465	77	74	3
Marchantia polymorpha (Asn:N)	X04465	75	72	3
Marchantia polymorpha (Pro:P)	X04465	72	69	3
Marchantia polymorpha (Pro:P)	X04465	77	74	3
Marchantia polymorpha (Gln:Q)	X04465	75	72	3
Marchantia polymorpha (Arg:R)	X04465	77	74	3
Marchantia polymorpha (Arg:R)	X04465	77	74	3
Marchantia polymorpha (Arg:R)	X04465	75	72	3
Marchantia polymorpha (Thr:T)	X04465	75	72	3
Marchantia polymorpha (Thr:T)	X04465	76	73	3
Marchantia polymorpha (Val:V)	X04465	75	72	3
Marchantia polymorpha (Val:V)	M20972	75	72	3
Marchantia polymorpha (Trp:W)	X04465	77	74	3
Medicago sativa (His:H)	AY029748	77	75	2
Medicago truncatula (Asp:D)	AC093544	77	74	3
Medicago truncatula (Met:M)	AC093544	76	73	3
Medicago truncatula (Pro:P)	AC093544	77	74	3
Medicago truncatula (Arg:R)	AC093544	75	72	3
Medicago truncatula (Thr:T)	AC093544	75	72	3
Medicago truncatula (Trp:W)	AC093544	77	74	3
Mesostigma viride (Lys:K)	AF166114	75	72	3
Nephroselmis olivacea (Ala:A)	AF137379	74	71	3
Nicotiana tabacum (Cys:C)	Z00044	75	72	3
Nicotiana tabacum (Asp:D)	Z00044	77	74	3

Nicotiana tabacum (Glu:E)	Z00044	76	73	3
Nicotiana tabacum (Phe:F)	Z00044	76	73	3
Nicotiana tabacum (Gly:G)	Z00044	74	71	3
Nicotiana tabacum (His:H)	Z00044	77	75	2
Nicotiana tabacum (Met:M)	Z00044	76	73	3
Nicotiana tabacum (Asn:N)	Z00044	75	72	3
Nicotiana tabacum (Pro:P)	Z00044	77	74	3
Nicotiana tabacum (Gln:Q)	Z00044	75	72	3
Nicotiana tabacum (Thr:T)	Z00044	75	72	3
Nicotiana tabacum (Trp:W)	Z00044	77	74	3
Nicotiana tabacum (Tyr:Y)	X00360	76	76	0
Nicotiana tabacum (Tyr:Y)	X00361	76	76	0
Parodia erinacea (Thr:T)	AY064336	76	73	3
Pelargonium zonale (Arg:R)	X01120	77	74	3
Phaseolus vulgaris (Phe:F)	K00336	76	76	0
Pisum sativum (Phe:F)	X04551	76	75	1
Pisum sativum (Gly:G)	X05394	74	71	3
Pisum sativum (Pro:P)	X05395	77	74	3
Pisum sativum (Arg:R)	M16863	77	74	3
Pisum sativum (Val:V)	X55033	75	72	3
Pisum sativum (Trp:W)	X05395	77	74	3
Pisum sativum (Trp:W)	X05395	77	74	3
Ptychosperma burretianum (Glu:E)	AF449169	76	73	3
Scenedesmus obliquus (Phe:F)	M25610	76	75	1
Scenedesmus obliquus (Met:M)	M25611	77	76	1
Scenedesmus obliquus (Tyr:Y)	X02224	76	75	1
Sinapis alba (His:H)	X17331	76	73	3
Sinapis alba (Gln:Q)	X13558	75	72	3
Spinacia oleracea (Cys:C)	AJ400848	74	71	3
Spinacia oleracea (Asp:D)	AJ400848	77	74	3
Spinacia oleracea (Glu:E)	AJ400848	76	73	3
Spinacia oleracea (Phe:F)	X02686	76	76	0
Spinacia oleracea (His:H)	AJ400848	77	75	2
Spinacia oleracea (Ile:I)	K01839	77	74	3
Spinacia oleracea (Ile:I)	K00222	75	72	3
Spinacia oleracea (Ile:I)	K00222	75	75	0
Spinacia oleracea (Ile:I)	K00222	75	72	3
Spinacia oleracea (Ile:I)	K02848	77	76	1
Spinacia oleracea (Met:M)	AJ400848	76	73	3
Spinacia oleracea (Pro:P)	K00358	77	74	3
Spinacia oleracea (Pro:P)	AJ400848	77	74	3
Spinacia oleracea (Arg:R)	AJ400848	75	72	3
Spinacia oleracea (Thr:T)	AJ400848	75	72	3
Spinacia oleracea (Thr:T)	K00281	75	75	0
Spinacia oleracea (Thr:T)	AJ400848	76	73	3
Spinacia oleracea (Val:V)	AJ400848	75	72	3
Spinacia oleracea (Val:V)	K00247	77	75	2
Spinacia oleracea (Val:V)	K00247	77	74	3

Spinacia oleracea (Trp:W)	K00262	77	75	2
Spirodela punctata (Arg:R)	X00764	75	72	3
Triticum aestivum (Met:M)	X02560	74	71	3
Triticum aestivum (Trp:W)	K02003	77	77	0
Triticum aestivum (Gly:G)	X00756	76	75	1
Vicia faba (Glu:E)	X00682	76	73	3
Vicia faba (Phe:F)	X51471	76	73	3
Zea mays (Cys:C)	X86563	74	71	3
Zea mays (Trp:W)	X86563	77	74	3

Eukaryotic Nuclear	**Accessions**	**Length**	**AUGC**	**Other**
Arabidopsis thaliana (Asp:D)	AC016041	75	72	3
Arabidopsis thaliana (Phe:F)	AC011665	76	73	3
Arabidopsis thaliana (Lys:K)	AC026234	76	76	0
Arabidopsis thaliana (Pro:P)	NM_105549	75	75	0
Arabidopsis thaliana (Pro:P)	NM_105549	75	72	3
Arabidopsis thaliana (Pro:P)	AC018907	75	72	3
Arabidopsis thaliana (Arg:R)	AB019236	76	76	0
Arabidopsis thaliana (Val:V)	AC025417	77	74	3
Bombyx mori (Ala:A)	M23363	76	73	3
Bombyx mori (Glu:E)	X03602	75	72	3
Bombyx mori (Gly:G)	K00206	74	74	0
Bos taurus (Asp:D)	K00175	75	72	3
Bos taurus (Phe:F)	K00352	76	76	0
Bos taurus (Arg:R)	V00134	76	76	0
Bos taurus (Arg:R)	X04541	76	76	0
Bos taurus (Thr:T)	M26109	76	76	0
Bos taurus (Trp:W)	M10543	75	75	0
Bos taurus (Tyr:Y)	M26210	76	76	0
Caenorhabditis elegans (Asp:D)	U41014	75	72	3
Caenorhabditis elegans (Lys:K)	AF040661	76	73	3
Caenorhabditis elegans (Pro:P)	AC024859	75	72	3
Caenorhabditis elegans (Trp:W)	U70846	75	72	3
Dictyostelium discoideum (Glu:E)	AF037042	75	72	3
Dictyostelium discoideum (Val:V)	AF067200	77	74	3
Dictyostelium discoideum (Val:V)	X03499	77	74	3
Drosophila melanogaster (Ala:A)	AC009461	76	73	3
Drosophila melanogaster (Asp:D)	NG_000295	75	72	3
Drosophila melanogaster (Glu:E)	V00238	75	72	3
Drosophila melanogaster (Glu:E)	AC010564	75	72	3
Drosophila melanogaster (Glu:E)	K00193	75	75	0
Drosophila melanogaster (Glu:E)	NG_000161	75	72	3
Drosophila melanogaster (Phe:F)	AC023722	76	73	3
Drosophila melanogaster (Phe:F)	K00349	76	76	0
Drosophila melanogaster (Gly:G)	NG_000194	74	71	3
Drosophila melanogaster (Gly:G)	X07778	74	71	3
Drosophila melanogaster (His:H)	AC099014	75	72	3
Drosophila melanogaster (His:H)	K00215	75	75	0

Drosophila melanogaster (Ile:I)	NG_000454	77	74	3
Drosophila melanogaster (Lys:K)	AC008257	76	73	3
Drosophila melanogaster (Lys:K)	AC008257	76	75	1
Drosophila melanogaster (Lys:K)	K01859	76	75	1
Drosophila melanogaster (Met:M)	K00462	75	72	3
Drosophila melanogaster (Asn:N)	AC008257	77	74	3
Drosophila melanogaster (Pro:P)	AC018491	75	72	3
Drosophila melanogaster (Pro:P)	AE003723	75	72	3
Drosophila melanogaster (Arg:R)	AC008257	76	73	3
Drosophila melanogaster (Arg:R)	AC021639	76	73	3
Drosophila melanogaster (Thr:T)	AC097445	77	74	3
Drosophila melanogaster (Val:V)	AC009461	76	73	3
Drosophila melanogaster (Val:V)	AC010713	76	76	0
Drosophila melanogaster (Val:V)	AC009461	76	73	3
Drosophila melanogaster (Val:V)	M25880	76	76	0
Drosophila melanogaster (Val:V)	AC091207	76	74	2
Drosophila melanogaster (Tyr:Y)	M26124	76	76	0
Gallus gallus (Lys:K)	J00881	76	73	3
Homo sapiens (Ala:A)	AC013472	76	76	0
Homo sapiens (Ala:A)	AL121936	76	76	0
Homo sapiens (Ala:A)	AL121932	75	72	3
Homo sapiens (Glu:E)	J00309	75	72	3
Homo sapiens (Glu:E)	AL355149	75	72	3
Homo sapiens (Phe:F)	AL662890	76	73	3
Homo sapiens (Gly:G)	K00208	74	74	0
Homo sapiens (Gly:G)	K00209	74	74	0
Homo sapiens (Gly:G)	AL355149	74	71	3
Homo sapiens (Gly:G)	K00208	74	74	0
Homo sapiens (His:H)	X01553	76	76	0
Homo sapiens (His:H)	U43279	75	72	3
Homo sapiens (His:H)	U43279	75	74	1
Homo sapiens (His:H)	X01553	75	75	0
Homo sapiens (Ile:I)	AL121934	77	77	0
Homo sapiens (Lys:K)	U00939	76	76	0
Homo sapiens (Asn:N)	AL356957	77	74	3
Homo sapiens (Asn:N)	K01921	77	74	3
Homo sapiens (Asn:N)	X15813	75	74	1
Homo sapiens (Pro:P)	AC024952	75	72	3
Homo sapiens (Pro:P)	AC008443	75	72	3
Homo sapiens (Gln:Q)	K01921	77	74	3
Homo sapiens (Gln:Q)	X15814	75	74	1
Homo sapiens (Gln:Q)	X15813	75	74	1
Homo sapiens (Arg:R)	AJ333675	76	76	0
Homo sapiens (Arg:R)	AL121936	76	76	0
Homo sapiens (Arg:R)	AC083880	76	76	0
Homo sapiens (Thr:T)	AL163636	76	73	3
Homo sapiens (Val:V)	AC008443	76	76	0
Homo sapiens (Val:V)	AC008443	76	73	3

Homo sapiens (Val:V)	AL031229	76	73	3
Homo sapiens (Val:V)	AC008443	76	73	3
Homo sapiens (Val:V)	AC008443	76	73	3
Homo sapiens (Val:V)	AC005783	76	73	3
Homo sapiens (Tyr:Y)	X04779	76	76	0
Hordeum vulgare (Glu:E)	X06283	75	75	0
Hordeum vulgare (Glu:E)	X06283	76	76	0
Hordeum vulgare (Glu:E)	X06378	76	76	0
Hordeum vulgare (Glu:E)	X06284	76	76	0
Hordeum vulgare (Gln:Q)	X06376	75	73	2
Lupinus luteus (Glu:E)	M23387	76	76	0
Lupinus luteus (Phe:F)	K00345	76	76	0
Lupinus luteus (Gly:G)	X05493	74	74	0
Lupinus luteus (His:H)	M16065	75	75	0
Lupinus luteus (Ile:I)	X06459	77	77	0
Lupinus luteus (Asn:N)	X07526	76	76	0
Lupinus luteus (Val:V)	X05082	76	76	0
Lupinus luteus (Val:V)	X05082	77	74	3
Mus musculus (Glu:E)	X00229	75	72	3
Mus musculus (Glu:E)	X00229	75	75	0
Mus musculus (Gly:G)	AC069308	74	71	3
Mus musculus (His:H)	J00642	75	72	3
Mus musculus (Ile:I)	AL589879	77	74	3
Mus musculus (Met:M)	X04525	75	75	0
Mus musculus (Asn:N)	AY050218	77	74	3
Mus musculus (Pro:P)	K00360	75	74	1
Mus musculus (Gln:Q)	AC092498	75	72	3
Mus musculus (Gln:Q)	M16252	75	75	0
Neurospora crassa (Phe:F)	X02710	76	74	2
Oryctolagus cuniculus (Asp:D)	K00176	76	76	0
Oryctolagus cuniculus (Lys:K)	K00289	76	76	0
Oryctolagus cuniculus (Met:M)	X68632	76	76	0
Oryza sativa (Cys:C)	AC092750	74	71	3
Oryza sativa (Phe:F)	AC092750	76	73	3
Oryza sativa (Gly:G)	AC092750	74	71	3
Oryza sativa (Ile:I)	AC099402	77	74	3
Oryza sativa (Met:M)	AC092750	76	73	3
Oryza sativa (Met:M)	AC092750	76	73	3
Oryza sativa (Asn:N)	AC099402	75	72	3
Oryza sativa (Arg:R)	AC099402	77	74	3
Oryza sativa (Thr:T)	AC092750	75	72	3
Oryza sativa (Thr:T)	AC092750	76	73	3
Oryza sativa (Val:V)	AC099402	75	72	3
Pichia jadinii (Ile:I)	K01061	77	77	0
Pichia jadinii (Pro:P)	K00357	75	75	0
Pichia jadinii (Tyr:Y)	M24830	78	78	0
Rattus norvegicus (Asp:D)	K00444	75	72	3
Rattus norvegicus (Asp:D)	K03129	75	72	3

Rattus norvegicus (Asp:D)	V01269	75	75	0
Rattus norvegicus (Glu:E)	V01272	75	72	3
Rattus norvegicus (Glu:E)	K00446	75	72	3
Rattus norvegicus (Glu:E)	K00446	75	72	3
Rattus norvegicus (Glu:E)	K00195	75	75	0
Rattus norvegicus (Phe:F)	M22764	77	74	3
Rattus norvegicus (Gly:G)	V01272	75	72	3
Rattus norvegicus (Gly:G)	X00706	75	72	3
Rattus norvegicus (Lys:K)	X04545	76	76	0
Rattus norvegicus (Asn:N)	K00166	77	76	1
Rattus norvegicus (Pro:P)	K01637	75	72	3
Rattus norvegicus (Gln:Q)	V01265	75	75	0
Rattus norvegicus (Val:V)	M34549	76	75	1
Saccharomyces cerevisiae (Cys:C)	M34549	75	72	3
Saccharomyces cerevisiae (Cys:C)	X01939	75	75	0
Saccharomyces cerevisiae (Asp:D)	X90518	75	72	3
Saccharomyces cerevisiae (Asp:D)	M25168	75	75	0
Saccharomyces cerevisiae (Glu:E)	U51030	75	72	3
Saccharomyces cerevisiae (Glu:E)	U18778	75	72	3
Saccharomyces cerevisiae (Glu:E)	K00191	75	75	0
Saccharomyces cerevisiae (Phe:F)	M10263	76	73	3
Saccharomyces cerevisiae (Phe:F)	M14867	76	76	0
Saccharomyces cerevisiae (Gly:G)	K00204	73	73	0
Saccharomyces cerevisiae (Gly:G)	U18779	74	71	3
Saccharomyces cerevisiae (Gly:G)	Z71561	75	74	1
Saccharomyces cerevisiae (Gly:G)	Z71561	75	75	0
Saccharomyces cerevisiae (His:H)	M26097	75	74	1
Saccharomyces cerevisiae (Ile:I)	U18922	77	74	3
Saccharomyces cerevisiae (Ile:I)	X69098	76	73	3
Saccharomyces cerevisiae (Lys:K)	K00286	76	76	0
Saccharomyces cerevisiae (Lys:K)	U18530	76	73	3
Saccharomyces cerevisiae (Lys:K)	K00287	76	73	3
Saccharomyces cerevisiae (Met:M)	J01372	76	76	0
Saccharomyces cerevisiae (Met:M)	M10268	76	76	0
Saccharomyces cerevisiae (Asn:N)	M26099	77	77	0
Saccharomyces cerevisiae (Pro:P)	M26096	75	75	0
Saccharomyces cerevisiae (Gln:Q)	X66375	75	71	4
Saccharomyces cerevisiae (Gln:Q)	U18796	75	72	3
Saccharomyces cerevisiae (Arg:R)	U18917	76	76	0
Saccharomyces cerevisiae (Arg:R)	L47993	75	72	3
Saccharomyces cerevisiae (Arg:R)	U18530	75	72	3
Saccharomyces cerevisiae (Arg:R)	K00158	75	75	0
Saccharomyces cerevisiae (Arg:R)	K00159	75	75	0
Saccharomyces cerevisiae (Thr:T)	K00279	76	76	0
Saccharomyces cerevisiae (Val:V)	Z75085	77	74	3
Saccharomyces cerevisiae (Val:V)	K00249	76	76	0
Saccharomyces cerevisiae (Val:V)	Z47814	77	77	0
Saccharomyces cerevisiae (Trp:W)	M35060	75	75	0

Saccharomyces cerevisiae (Tyr:Y)	M10266	78	78	0
Saccharomyces pastorianus (Phe:F)	X00655	76	76	0
Schizosaccharomyces pombe (Asp:D)	AL590457	74	71	3
Schizosaccharomyces pombe (Glu:E)	AL121794	75	72	3
Schizosaccharomyces pombe (Phe:F)	Z97208	76	73	3
Schizosaccharomyces pombe (Phe:F)	K00344	76	72	4
Schizosaccharomyces pombe (His:H)	AL031825	75	72	3
Schizosaccharomyces pombe (Lys:K)	Z97185	78	75	3
Schizosaccharomyces pombe (Arg:R)	X00239	76	73	3
Schizosaccharomyces pombe (Arg:R)	AL590457	76	73	3
Schizosaccharomyces pombe (Arg:R)	AL590457	76	75	1
Schizosaccharomyces pombe (Tyr:Y)	K00273	77	77	0
Sorghum bicolor (Gly:G)	AF466201	74	71	3
Tetrahymena pyriformis (Asn:N)	X16643	77	74	3
Tetrahymena thermophila (Gln:Q)	M35401	75	74	1
Tetrahymena thermophila (Gln:Q)	M11464	75	75	0
Tetrahymena thermophila (Gln:Q)	M35400	75	75	0
Trypanosoma brucei (Lys:K)	AF047724	76	72	4
Trypanosoma brucei (Val:V)	X16590	76	73	3
Trypanosoma brucei rhodesiense (Gln:Q)	X16590	75	72	3
Xenopus laevis (Ala:A)	Y00430	75	72	3
Xenopus laevis (Asp:D)	X04460	75	75	0
Xenopus laevis (Phe:F)	K02849	76	76	0
Xenopus laevis (Lys:K)	Y00163	76	73	3
Xenopus laevis (Val:V)	X04819	76	73	3

5S Ribosomal RNA (5S rRNA)	**Accessions**	**Length**	**AUGC**	**Other**
Archaea				
Haloarcula marismortui	AF034620	122	122	0
Haloferax mediterranei	X14441	123	123	0
Methanocaldococcus jannaschii	U67518	122	122	0
Methanolobus tindarius	M34910	128	128	0
Methanothermococcus thermolithotrophicus	M34911	120	120	0
Methanothermus fervidus	M26976	124	124	0
Pyrococcus woesei	X15329	124	124	0
Pyrodictium occultum	M21086	130	130	0
Sulfolobus acidocaldarius	V01286	126	125	1
Sulfolobus solfataricus	X01588	126	126	0
Thermococcus celer	X07692	127	127	0
Thermoplasma acidophilum	M32297	123	123	0
Bacteria	**Accessions**	**Length**	**AUGC**	**Other**
Acidithiobacillus ferrooxidans	M11542	120	120	0
Agrobacterium tumefaciens	X02627	120	120	0
Arthrobacter globiformis	M16173	123	122	1
Arthrobacter globiformis	X08002	122	121	1
Arthrobacter oxydans	X08000	122	121	1

Bacillus subtilis	D11460	118	115	3
Campylobacter jejuni	AL139076	121	118	3
Deinococcus radiodurans	AE002087	124	124	0
Delftia acidovorans	AJ131594	117	115	2
Escherichia coli	V00336	120	120	0
Geobacillus stearothermophilus	M10816	119	117	2
Geobacillus stearothermophilus	AJ251080	117	117	0
Geobacillus stearothermophilus	M24839	119	119	0
Geobacillus stearothermophilus	M25591	117	117	0
Haemophilus influenzae	U32688	120	115	5
Micrococcus luteus	K02682	120	119	1
Mycoplasma genitalium	U39694	118	118	0
Planctomyces brasiliensis	M35168	113	110	3
Pseudomonas aeruginosa	K02353	120	120	0
Pseudomonas stutzeri	M34776	119	118	1
Rhodobacter capsulatus	X04585	119	118	1
Spiroplasma melliferum	X06098	109	107	2
Sporosarcina pasteurii	X02024	119	117	2
Staphylococcus aureus	L36472	118	116	2
Synechococcus sp. PCC 6301	X00757	121	120	1
Thermus aquaticus	X01590	123	123	0
Thermus sp.	M16532	121	120	1
Thermus thermophilus	V01415	121	120	1

Eukaryotic Chloroplast	**Accessions**	**Length**	**AUGC**	**Other**
Chlamydomonas reinhardtii	BK000554	121	121	0
Euglena gracilis	K02483	118	118	0
Marchantia polymorpha	X00666	119	119	0
Zea mays	M19943	121	121	0

Eukaryotic Mitochondrion	**Accessions**	**Length**	**AUGC**	**Other**
Reclinomonas americana	U59762	115	114	1

Eukaryotic Nuclear	**Accessions**	**Length**	**AUGC**	**Other**
Acanthamoeba castellanii	V00003	119	119	0
Acheta domesticus	M16074	120	120	0
Amoebidium parasiticum	M36306	119	119	0
Ascobolus immersus	X99087	119	119	0
Asterias vulgaris	X00992	120	120	0
Aurelia aurita	X00991	119	118	1
Blastocladiella simplex	X01543	118	118	0
Blepharisma japonicum	J01851	120	120	0
Bos taurus	X57170	120	120	0
Branchiostoma belcheri	X13034	120	120	0
Candida albicans	X00868	121	121	0
Chlamydomonas reinhardtii	X02706	122	122	0
Christiansenia pallida	M58383	120	120	0
Crithidia fasciculata	V00149	120	120	0

Crypthecodinium cohnii	M25115	122	122	0
Cryptococcus neoformans var. neoformans	L14753	118	118	0
Cyanophora paradoxa	M33029	119	119	0
Diatoma tenue	D00058	118	118	0
Drosophila melanogaster	M25016	120	120	0
Dugesia japonica	X01551	120	120	0
Emplectonema gracile	X00021	120	119	1
Enchytraeus albidus	X03911	120	120	0
Equisetum arvense	X00377	120	120	0
Euglena gracilis	X01484	123	121	2
Exobasidium vaccinii	X00069	118	118	0
Globodera pallida	L28955	120	120	0
Gracilaria compressa	X00999	121	121	0
Homo sapiens	Z75742	119	119	0
Hyphodontia paradoxa	X73890	118	118	0
Mesocricetus auratus	J00063	121	121	0
Mortierella formosensis	M36312	120	120	0
Octopus vulgaris	X06835	120	120	0
Oryza sativa	M18171	119	119	0
Phaseolus vulgaris	X06843	120	120	0
Physarum polycephalum	X02036	120	120	0
Plagiomnium trichomanes	X01619	119	119	0
Plasmodium falciparum	AF239766	122	119	3
Pneumocystis carinii	M28193	120	120	0
Pseudocentrotus depressus	X04307	120	120	0
Saccharomyces cerevisiae	X67579	118	118	0
Schizochytrium aggregatum	X06104	119	119	0
Schizosaccharomyces pombe	K00570	119	119	0
Spirogyra sp.	M10438	120	120	0
Tetrahymena thermophila	X00475	120	120	0
Xenopus laevis	X05089	120	120	0

16S Ribosomal RNA (16S rRNA)	**Accessions**	**Length**	**AUGC**	**Other**
Archaea				
Aeropyrum pernix	AP000062	1504	1501	3
Archaeoglobus fulgidus	X05567	1493	1492	1
Haloarcula marismortui rrnA	X61688	1473	1472	1
Haloarcula marismortui rrnB	X61689	1473	1472	1
Halobacterium sp.	AE005128	1473	1473	0
Haloferax volcanii	K00421	1474	1474	0
Methanobacterium formicicum	M36508	1477	1476	1
Methanococcus jannaschii	U67517	1478	1478	0
Methanococcus vannielii	M36507	1468	1466	2
Methanospirillum hungatei	M60880	1467	1465	2
Methanothermobacter thermautotrophicus	AE000930	1481	1479	2
Natronobacterium innermongoliae	AF009601	1473	1472	1
Natronorubrum bangense	Y14028	1476	1475	1

Pyrococcus abyssi	AJ248283	1512	1512	0
Pyrococcus furiosus	U20163	1496	1495	1
Pyrococcus horikoshii	AP000001	1500	1500	0
Pyrodictium occultum	M21087	1498	1494	4
Sulfolobus acidocaldarius	D14876	1495	1492	3
Sulfolobus P2	**NPA**	1497	1495	2
Sulfolobus solfataricus	X03235	1495	1492	3
Thermococcus celer	M21529	1487	1486	1
Thermoplasma acidophilum	AL445067	1473	1471	2
Thermoproteus tenax	M35966	1505	1503	2

Bacteria	**Accessions**	**Length**	**AUGC**	**Other**
Acidobacterium capsulatum	D26171	1498	1418	80
Acinetobacter calcoaceticus	M34139	1537	1506	31
Actinomyces israelii	X82450	1550	1439	111
Aeromonas hydrophila	X60407	1544	1502	42
Agrobacterium tumefaciens	M11223	1489	1489	0
Allochromatium vinosum	M26629	1528	1483	45
Anabaena sp.	X59559	1489	1489	0
Aquifex aeolicus	AE000709	1588	1587	1
Aquifex pyrophilus	M83548	1582	1564	18
Arthrobacter globiformis	M23411	1531	1531	0
Azorhizobium caulinodans	D11342	1482	1467	15
Bacillus anthracis	X55059	1552	1422	130
Bacillus cereus	X55060	1551	1440	111
Bacillus halodurans	AB013373	1554	1553	1
Bacillus subtilis	K00637	1552	1552	0
Bacteroides fragilis	M61006	1537	1536	1
Bartonella bacilliformis	Z11683	1488	1399	89
Bartonella henselae	M73229	1487	1412	75
Bartonella quintana	M11927	1493	1492	1
Beggiatoa sp. 1401-13	L40997	1522	1459	63
Bordetella bronchiseptica	U04948	1532	1532	0
Bordetella parapertussis	U04949	1531	1458	73
Bordetella pertussis	U04950	1531	1458	73
Borrelia burgdorferi	M88329	1537	1537	0
Borrelia hermsii	U42292	1536	1523	13
Brachyspira hyodysenteriae	U23035	1515	1463	52
Bradyrhizobium japonicum	Z35330	1490	1490	0
Brevinema andersonii	L31543	1530	1443	87
Brucella melitensis	L26166	1489	1402	87
Buchnera sp. APS	AP000398	1548	1460	88
Burkholderia mallei	S55008	1533	1340	193
Burkholderia sp.	U37342	1535	1513	22
Campylobacter jejuni	Z29326	1515	1513	2
Campylobacter sputorum	X67775	1744	1694	50
Chlamydia muridarum	AE002280	1554	1550	4
Chlamydia trachomatis	U68443	1554	1554	0

Chlamydophila pneumoniae	L06108	1554	1554	0
Chlamydophila psittaci	U68447	1552	1552	0
Chlorobium vibrioforme	M62791	1507	1503	4
Chlorogloeopsis sp. PCC 7518	X68780	1490	1482	8
Chromohalobacter marismortui	X87222	1538	1474	64
Citrobacter freundii	M59291	1542	1457	85
Clostridium botulinum F	L37593	1512	1451	61
Clostridium difficile	X73450	1503	1462	41
Clostridium innocuum	M23732	1543	1538	5
Clostridium perfringens	M69264	1514	1513	1
Clostridium tetani	X74770	1517	1508	9
Comamonas testosteroni	M11224	1536	1536	0
Corynebacterium diphtheriae	X84248	1520	1482	38
Coxiella burnetii	M21291	1541	1471	70
Cristispira CP1	U42638	1528	1491	37
Deferribacter thermophilus	U75602	1559	1551	8
Deinococcus radiodurans	M21413	1504	1502	2
Desulfovibrio desulfuricans	M34113	1551	1550	1
Dichelobacter nodosus	M35016	1533	1529	4
Edwardsiella tarda	AF015259	1542	1428	114
Enterococcus faecalis	Y18293	1561	1449	112
Enterococcus faecium	AF070223	1538	1510	28
environ.Eubacteria clone W15	NPA	1521	1433	88
Epulopiscium sp.	M99572	1511	1433	78
Erysipelothrix rhusiopathiae	M23728	1539	1480	59
Escherichia coli	J01695	1542	1542	0
Eubacterium brachy	Z36272	1520	1500	20
Francisella tularensis	Z21931	1525	1521	4
Frankia sp.	M55343	1512	1512	0
Fusobacterium necrophorum	X74408	1507	1499	8
Fusobacterium nucleatum subsp. nucleatum	M58683	1522	1494	28
Gemmata obscuriglobus	X56305	1497	1452	45
Geotoga subterranea	L10659	1530	1528	2
Gluconacetobacter liquefaciens	X75617	1494	1486	8
Haemobartonella felis	U95297	1487	1377	110
Haemophilus influenzae	X87977	1539	1441	98
Haemophilus influenzae (operons A-F)	U32741	1539	1539	0
Halomonas halodenitrificans	L04942	1538	1526	12
Helicobacter pylori	M88157	1503	1444	59
Heliobacterium chlorum	M11212	1526	1512	14
Holophaga foetida	X77215	1540	1500	40
Isosphaera pallida	X64372	1518	1438	80
Klebsiella pneumoniae	X80684	1541	1443	98
Lactobacillus acidophilus	M58802	1568	1528	40
Lactococcus lactis subsp. lactis	AE006456	1551	1551	0
Legionella pneumophila	M59157	1543	1501	42
Leptonema illini	M88719	1528	1526	2
Leptospira interrogans	X17547	1508	1508	0

Leptospirillum ferriphilum	AF356830	1553	1522	31
Listeria monocytogenes	M58822	1552	1503	49
Mesorhizobium loti	AP003001	1486	1482	4
Methylobacterium sp.	Z23160	1480	1437	43
Methylococcus capsulatus	X72771	1534	1472	62
Micrococcus luteus	M38242	1524	1492	32
Microcystis aeruginosa	U03402	1489	1411	78
Mycobacterium avium	X52918	1533	1460	73
Mycobacterium leprae	X56657	1548	1548	0
Mycobacterium tuberculosis	X52917	1534	1464	70
Mycoplasma capricolum	X00921	1525	1523	2
Mycoplasma gallisepticum	M22441	1519	1519	0
Mycoplasma genitalium	U39694	1519	1519	0
Mycoplasma hyopneumoniae	Y00149	1537	1537	0
Mycoplasma mycoides	M23943	1524	1399	125
Mycoplasma pneumoniae	M29061	1516	1461	55
Myxococcus xanthus	M34114	1540	1539	1
Neisseria gonorrhoeae	X07714	1544	1544	0
Neisseria meningitidis	AE002364	1544	1544	0
Nocardia asteroides	X80606	1517	1469	48
Oscillatoria agardhii	X84811	1490	1463	27
Pasteurella multocida	M35018	1542	1500	42
Petrotoga miotherma	L10657	1529	1327	202
Pirellula marina	X62912	1472	1472	0
Pirellula staleyi	M34126	1525	1521	4
Planctomycetaceae Schlesner 670	X81948	1519	1490	29
Plesiomonas shigelloides	X74688	1542	1454	88
Pleurocapsa sp.	X78681	1489	1461	28
Porphyromonas gingivalis	L16492	1531	1467	64
Proteus vulgaris	X07652	1543	1543	0
Pseudomonas aeruginosa	M34133	1536	1484	52
Pseudomonas putida	D84020	1537	1527	10
Pseudomonas sp.	U37339	1537	1469	68
Psychrobacter pacificensis	AB016054	1536	1536	0
Rhodobium orientis	D30792	1486	1408	78
Rhodoblastus acidophilus	M34128	1491	1454	37
Rhodococcus erythropolis	AF001265	1519	1519	0
Rickettsia bellii	U11014	1502	1501	1
Rickettsia prowazekii	M21789	1502	1502	0
Rickettsia rickettsii	L36217	1499	1440	59
Salmonella typhimurium	X80681	1540	1533	7
Serratia marcescens	M59160	1542	1511	31
Shewanella putrefaciens	X81623	1543	1534	9
Shigella dysenteriae	X96966	1540	1487	53
Spirochaeta aurantia	M57740	1560	1520	40
Staphylococcus aureus	L36472	1555	1555	0
Streptobacillus moniliformis	Z35305	1521	1485	36
Streptococcus mutans	X58303	1546	1307	239

Streptococcus pneumoniae	X58312	1550	1290	260
Streptococcus pyogenes	X59029	1549	1278	271
Streptomyces acidiscabies	D63865	1530	1530	0
Streptomyces albidoflavus	Z76676	1527	1475	52
Streptomyces albus	X53163	1525	1240	285
Streptomyces ambofaciens	M27245	1529	1528	1
Streptomyces bikiniensis	X79851	1525	1517	8
Streptomyces bluensis	X79324	1528	1520	8
Streptomyces bottropensis	D63868	1531	1531	0
Streptomyces brasiliensis	X53162	1522	1238	284
Streptomyces caelestis	X80824	1526	1518	8
Streptomyces coelicolor	Y00411	1529	1528	1
Streptomyces diastaticus	X53161	1521	1189	332
Streptomyces diastatochromogenes	D63867	1531	1531	0
Streptomyces espinosus	X80826	1526	1518	8
Streptomyces eurythermus	D63870	1531	1531	0
Streptomyces felleus	Z76681	1527	1475	52
Streptomyces galbus	X79325	1525	1517	8
Streptomyces glaucescens	X79322	1527	1519	8
Streptomyces gougerotii	Z76687	1527	1476	51
Streptomyces griseus	X61478	1528	1528	0
Streptomyces hygroscopicus	X79853	1525	1517	8
Streptomyces intermedius	Z76686	1525	1474	51
Streptomyces lavendulae	X53173	1520	1183	337
Streptomyces limosus	Z76679	1527	1475	52
Streptomyces lincolnensis	X79854	1527	1519	8
Streptomyces macrosporus	Z68099	1529	1482	47
Streptomyces mashuensis	X79323	1526	1518	8
Streptomyces megasporus	Z68100	1535	1488	47
Streptomyces neyagawaensis	D63869	1531	1531	0
Streptomyces nodosus	AF114033	1528	1528	0
Streptomyces odorifer	Z76682	1527	1475	52
Streptomyces ornatus	X79326	1525	1517	8
Streptomyces pseudogriseolus	X80827	1524	1516	8
Streptomyces purpureus	X53170	1521	1185	336
Streptomyces rimosus	X62884	1530	1529	1
Streptomyces rutgersensis	Z76688	1527	1476	51
Streptomyces sampsonii	D63871	1531	1531	0
Streptomyces scabiei	D63862	1530	1530	0
Streptomyces setonii	D63872	1532	1532	0
Streptomyces sp.	D63866	1530	1530	0
Streptomyces subrutilus	X80825	1524	1516	8
Streptomyces tendae	D63873	1530	1530	0
Streptomyces thermodiastaticus	Z68101	1528	1483	45
Streptomyces thermolineatus	Z68097	1526	1481	45
Streptomyces thermoviolaceus	Z68096	1528	1483	45
Streptomyces thermovulgaris	Z68098	1531	1486	45
Synechococcus sp. PCC 6301	X03538	1488	1488	0

Synechocystis sp. PCC 6803	D64000	1489	1489	0
Thermomicrobium roseum	M34115	1527	1522	5
Thermotoga maritima	M21774	1562	1562	0
Thermus aquaticus	L09663	1519	1470	49
Thermus thermophilus	X07998	1518	1515	3
Treponema pallidum (rRNA A)	AE001204	1549	1549	0
Tropheryma whipplei	X99636	1522	1522	0
Ureaplasma urealyticum	AE002112	1545	1545	0
Vibrio cholerae	X76337	1544	1536	8
Vibrio parahaemolyticus	X56580	1546	1479	67
Xanthomonas albilineans	X95918	1545	1498	47
Xanthomonas campestris	**NPA**	1547	1484	63
Xylella fastidiosa	AE003861	1545	1545	0
Yersinia pestis	L37604	1542	1467	75
Yersinia pseudotuberculosis	Z21939	1543	1477	66

Eukaryotic Chloroplast	**Accessions**	**Length**	**AUGC**	**Other**
Apodanthes sp	**NPA**	1502	1502	0
Astasia longa	X14386	1520	1520	0
Babesia bovis	U06105	1486	1448	38
Chlamydomonas humicola	AF374186	1496	1325	171
Chlamydomonas reinhardtii	J01395	1474	1474	0
Chlorella vulgaris	AB001684	1494	1491	3
Corethron criophilum	**NPA**	1477	1477	0
Cryptomonas sp.	X56805	1493	1493	0
Cyanidium caldarium	X52985	1492	1492	0
Cyanophora paradoxa	X81840	1506	1504	2
Cynomorium coccineum	U67743	1480	1480	0
Cytinus ruber	U47845	1497	1497	0
Emiliania huxleyi	X82156	1483	1483	0
Euglena gracilis	X12890	1491	1491	0
Glaucocystis nostochinearum	X82496	1519	1512	7
Gloeochaete wittrockiana	X82495	1519	1512	7
Heterosigma akashiwo	M82860	1537	1537	0
Hydnora africana	U67745	1504	1504	0
Marchantia polymorpha	X04465	1496	1496	0
Mitrastema yamamotoi	U67742	1465	1465	0
Nicotiana tabacum	V00165	1486	1486	0
Palmaria palmata	Z18289	1486	1486	0
Pilostyles thurberi	U67741	1531	1464	67
Plasmodium falciparum (plastid-like)	X57167	1426	1426	0
Plasmodium vivax	AF040974	1447	949	498
Polytoma obtusum	AF374187	1503	1329	174
Polytoma oviforme	AF374188	1485	1311	174
Polytoma uvella	AF374189	1589	1586	3
Pylaiella littoralis	X14873	1505	1505	0
Ricinus communis	L37580	1489	1423	66
Skeletonema pseudocostatum	X82155	1486	1485	1

Toxoplasma gondii	U87145	1500	1500	0
Zea mays	Z00028	1490	1490	0

Eukaryotic Mitochondrion	**Accessions**	**Length**	**AUGC**	**Other**
Acanthamoeba castellanii	U03732	1560	1560	0
Afrixalus fornasini	**NPA**	931	904	27
Albinaria caerulea	X83390	759	759	0
Alligator mississippiensis	L28074	939	939	0
Amblysomus hottentotus	M95108	952	952	0
Anas platyrhynchos	L16770	985	985	0
Anopheles gambiae	L20934	800	800	0
Anopheles quadrimaculatus	L04272	794	794	0
Antilocapra americana	M55540	958	958	0
Apis mellifera	L06178	786	786	0
Artemia franciscana	X69067	712	712	0
Ascaris suum	X54253	701	701	0
Aspergillus nidulans	J01393	1437	1437	0
Asterina pectinifera	D16387	950	950	0
Balaenoptera musculus	X72204	972	972	0
Bos taurus	J01394	955	955	0
Bufo boreas boreas	**NPA**	932	911	21
Bufo peltocephalus	**NPA**	936	916	20
Caenorhabditis elegans	X54252	697	697	0
Cafeteria roenbergensis	AF193903	1662	1662	0
Ceratophrys sp.	**NPA**	936	908	28
Chlamydomonas eugametos	AF008237	1257	1257	0
Chlamydomonas reinhardtii	X54860	1200	1200	0
Chondrus crispus	Z30950	1376	1376	0
Chorthippus parallelus ESC	X95574	794	794	0
Chorthippus parallelus NOR	X95575	794	794	0
Chrysodidymus synuroideus	**NPA**	1611	1606	5
Chrysodidymus synuroideus mg	AF222718	1579	1579	0
Coscoroba coscoroba	S76216	983	983	0
Coturnix coturnix	X57245	971	971	0
Crithidia fasciculata	X02548	612	612	0
Crossostoma lacustre	M91245	951	951	0
Cygnus melancoryphus	S76217	989	989	0
Cyprinus carpio	X61010	951	951	0
Damaliscus pygargus	M86499	953	953	0
Daphnia pulex	Z15015	751	751	0
Dictyostelium discoideum	D16466	1551	1551	0
Didelphis virginiana	Z29573	951	951	0
Drosophila teissieri	X54011	790	790	0
Drosophila virilis	X05914	784	784	0
Drosophila yakuba	X03240	789	789	0
Eleutherodactylus coqui	NPA	946	875	71
Equus caballus	X79547	975	975	0
Farfantepenaeus notialis	X84357	858	858	0

Felis catus	U20753	960	960	0
Gallus gallus	X52392	976	976	0
Glycine max	M16859	1990	1990	0
Harpactes ardens	U94810	979	957	22
Harpochytrium sp. JEL94	AY182005	1368	1368	0
Herpetomonas megaseliae	U01006	547	547	0
Homo sapiens	J01415	954	954	0
Katharina tunicata	U09810	882	881	1
Latimeria chalumnae	Z21921	975	975	0
Leishmania tarentolae	M10126	670	670	0
Locusta migratoria	X80245	827	827	0
Loxodonta africana	U60182	961	961	0
Lumbricus terrestris	U24570	785	785	0
Lutreolina crassicaudata	U33494	953	953	0
Macropus giganteus	X86941	952	952	0
Magicicada tredecim	**NPA**	750	727	23
Marchantia polymorpha	M68292	1974	1974	0
Metridium senile	S75445	1138	1138	0
Monosiga brevicollis	AF538053	1596	1596	0
Mus musculus	J01420	956	956	0
Muscardinus avellanarius	X84384	956	956	0
Mytilus edulis	M83756	945	945	0
Nephroselmis olivacea	AF110138	1509	1509	0
Neurospora crassa	L33367	1876	1872	4
Ochromonas danica	AF287134	1563	1563	0
Oenothera berteriana	X61277	1900	1900	0
Oncorhynchus mykiss	L29771	944	944	0
Ornithorhynchus anatinus	U33498	944	944	0
Pan troglodytes	D38113	949	949	0
Paracentrotus lividus	J04815	878	878	0
Paramecium tetraurelia	K01751	1677	1675	2
Pedinomonas minor	AF116775	1178	1178	0
Penicillium chrysogenum	Z23072	1449	1449	0
Petromyzon marinus	U11880	900	900	0
Phalanger orientalis	U33496	950	950	0
Phascogale tapoatafa	U33497	957	957	0
Phoca vitulina	X63726	961	961	0
Physarum polycephalum	X75592	1861	1861	0
Phytophthora infestans	U17009	1503	1503	0
Pichia canadensis	D49702	1537	1537	0
Podospora anserina	X14734	1779	1779	0
Porphyra purpurea	AF114794	1407	1407	0
Protopterus dolloi	L42813	933	933	0
Prototheca wickerhamii	X15435	1675	1675	0
Puma concolor	U33495	960	960	0
Pylaiella littoralis	X14874	1519	1519	0
Rana catesbeiana	X12841	937	937	0
Rattus norvegicus	J01438	953	953	0

Reclinomonas americana	AF007261	1595	1595	0
Rhizopus stolonifer	**NPA**	1431	1430	1
Rhodomonas salina	AF288090	1483	1483	0
Saccharomyces cerevisiae	V00704	1686	1686	0
Salmo salar	U12143	948	946	2
Sceloporus undulatus	L28075	947	946	1
Schizosaccharomyces pombe	X15738	1398	1398	0
Scyliorhinus canicula	Y16067	960	960	0
Secale cereale	Z14059	1977	1977	0
Sphenodon punctatus	L28076	904	904	0
Spizellomyces punctatus	AF404303	1213	1213	0
Stenella coeruleoalba	X78169	974	974	0
Suillus sinuspaulianus	L47584	1987	1987	0
Tetrahymena pyriformis	M12714	1669	1669	0
Trachemys scripta	L28077	966	966	0
Triticum aestivum	Z14078	1957	1957	0
Trypanosoma brucei	X02547	621	621	0
Williopsis saturnus var. mrakii	X71392	1614	1610	4
Xenopus laevis	M27605	945	945	0
Zea mays	X00794	1962	1962	0

Eukaryotic Nuclear	**Accessions**	**Length**	**AUGC**	**Other**
Acanthamoeba castellanii	U07413	2290	2290	0
Agmasoma penaei	**NPA**	1284	1111	173
Ahnfeltia plicata	Z14139	1765	1765	0
Alexandrium fundyense	U09048	1801	1797	4
Amblyospora sp.	U68474	1394	1348	46
Ameson michaelis	L15741	1280	1275	5
Androctonus australis	X77908	1812	1812	0
Antonospora scoticae	AF024655	1392	1371	21
Artemia salina	X01723	1810	1810	0
Audouinella dasyae	L26181	1772	1771	1
Audouinella hermannii	AF026040	1771	1771	0
Aulacoseira ambigua	X85404	1847	1844	3
Babesia bigemina	X59604	1701	1701	0
Bacillidium sp.	AF104087	1413	1386	27
Balamuthia mandrillaris	AF019071	2017	2017	0
Balbiania investiens	AF132294	1772	1715	57
Bangia sp. (Northwest Territories/NWT)	AF043355	1830	1797	33
Bangiopsis subsimplex	AF168627	1792	1769	23
Batrachospermum gelatinosum	AF026045	1765	1765	0
Batrachospermum macrosporum	AF026048	1770	1770	0
Bonnemaisonia hamifera	L26182	1765	1764	1
Bostrychia moritziana	AF203893	1793	1793	0
Candida albicans	M60302	1787	1787	0
Ceramium rubrum	L26183	1778	1778	0
Chlorella luteoviridis	X73998	1800	1800	0
Chondrus crispus	Z14140	1778	1778	0

Compsopogon coeruleus	AF087124	1802	1728	74
Corallina officinalis	L26184	1786	1786	0
Crossodonthina koreana	Z36893	1811	1811	0
Cryptocercus punctulatus	**NPA**	2016	2016	0
Cryptococcus neoformans var. neoformans	L05428	1802	1802	0
Culicosporella lunata	AF027683	1343	1343	0
Cyanophora paradoxa	X68483	1807	1807	0
Cymatosira belgica	X85387	2317	2316	1
Cyrtohymena citrina	AF164135	1772	1772	0
Dixonielloa grisea	L26187	1793	1789	4
Drosophila melanogaster	M21017	1995	1995	0
Echinococcus granulosus	U27015	2394	2394	0
Edhazardia aedis	AF027684	1483	1448	35
Encephalitozoon cuniculi	X98467	1295	1295	0
Encephalitozoon hellem	AF118143	1314	1314	0
Encephalitozoon sp.	L16867	1295	1295	0
Endoreticulatus schubergi	L39109	1252	1252	0
Engelmanniella mobilis	AF164134	1773	1773	0
Enterocytozoonidae gen. sp.	AF201911	1318	1300	18
Erythrotrichia carnea	L26189	1796	1795	1
Euglypha rotunda	X77692	1811	1811	0
Euplotes aediculatus	M14590	1882	1882	0
Flabelliforma montana	AJ252962	1383	1115	268
Fragaria x ananassa	X15590	1804	1804	0
Gastrostyla steinei	AF164133	1771	1771	0
Gelidium vagum	L26190	1769	1769	0
Genicularia spirotaenia	**NPA**	1803	1803	0
Giardia ardeae	Z17210	1435	1435	0
Giardia intestinalis	X52949	1452	1452	0
Giardia muris	X65063	1432	1432	0
Glaucocystis nostochinearum	X70803	1819	1819	0
Gloeochaete wittrockiana	X81901	1801	1801	0
Glomus intraradices	X58725	1800	1739	61
Glugea atherinae	U15987	1357	1335	22
Glugea stephani	AF056015	1314	1165	149
Gracilariopsis sp.	M33639	1782	1782	0
Halymenia plana	U33133	1770	1770	0
Hexamita sp.	Z17224	1550	1550	0
Hildenbrandia rubra	L19345	1781	1780	1
Homo sapiens	K03432	1870	1870	0
Ichthyosporidium sp.	L39110	1360	1352	8
Janacekia debaisieuxi	AJ252950	1417	1417	0
Lecanora dispersa	**NPA**	1797	1644	153
Lilioceris lilii	**NPA**	1921	1882	39
Loma acerinae	AJ252951	1352	1352	0
Mastigamoeba balamuthi	L23799	2741	2724	17
Microgemma sp.	AJ252952	1356	1348	8
Microsporidium 57864	U90885	1245	1245	0

Mus musculus	X00686	1869	1869	0
Mytilus edulis	L24489	1826	1826	0
Naegleria gruberi	**NPA**	2019	2019	0
Nemalion helminthoides	L26196	1770	1769	1
Nemalionopsis shawii	AF506272	1793	1736	57
Nosema algerae	AF069063	1390	1390	0
Nosema apis	U97150	1242	1242	0
Nosema necatrix	U11051	1247	1247	0
Okanagana utahensis	U06478	1918	1917	1
Onychodromus quadricornutus	X53485	1771	1770	1
Ophiopholis aculeata	L28056	1801	1801	0
Oxytricha granulifera	AF164122	1774	1774	0
Oxytricha longa	AF164125	1770	1770	0
Palmaria palmata	Z14142	1771	1771	0
Paraurostyla weissei	AF164127	1771	1771	0
Paruroleptus lepisma	AF164132	1772	1772	0
Paulinella chromatophora	X81811	1817	1815	2
Placopecten magellanicus	X53899	1815	1815	0
Plasmodium falciparum (A gene)	M19172	2090	2090	0
Plasmodium falciparum (S gene)	M19173	2145	2145	0
Plasmodium vivax (A gene)	U07367	2063	2063	0
Plasmodium vivax (O gene)	U93095	2230	2229	1
Plasmodium vivax (S gene)	U07368	2147	2147	0
Pleistophora hippoglossoideos	AJ252953	1372	1372	0
Pleistophora sp.	U10342	1254	1232	22
Pleurotricha lanceolota	AF164128	1771	1771	0
Plocamiocolax pulvinata	U09618	1771	1771	0
Polydispyrenia simulii	AJ252960	1398	1291	107
Porphyra miniata	AF175540	1836	1801	35
Porphyridium aerugineum	L27635	1778	1762	16
Reticulitermes flavipes	**NPA**	1882	1882	0
Rhodella maculata	U21217	1784	1755	29
Rhodochaete parvula	AF139462	1790	1758	32
Rhodogorgon carriebowensis	AF006089	1841	1841	0
Rhodymenia leptophylla	U09621	1771	1771	0
Saccharomyces cerevisiae	U53879	1800	1800	0
Spraguea lophii	AF033197	1400	1398	2
Staurastrum sp. M752	X74752	1803	1803	0
Strongylocentrotus purpuratus	L28055	1796	1796	0
Stylonychia lemnae	AF164124	1770	1770	0
Stylonychia mytilus	AF164123	1771	1771	0
Thalassiosira eccentrica	X85396	1802	1800	2
Thelohania solenopsae	AF031538	1410	1382	28
Thelohania sp.	AF031537	1387	1130	257
Thorea hispida	AF506273	1867	1759	108
Thorea violacea	AF026042	1916	1916	0
Thorea violacea	AF506274	1868	1752	116
Toxoplasma gondii (P)	X75453	1793	1793	0

Trachipleistophora hominis	AJ002605	1362	1362	0
Tritrichomonas foetus	M81842	1571	1571	0
Uroleptus gallina	AF164130	1775	1775	0
Uroleptus pisces	AF164131	1772	1772	0
Urostyla grandis	AF164129	1768	1768	0
Vairimorpha necatrix	Y00266	1244	1244	0
Vairimorpha sp. Argentina	AF031539	1331	1252	79
Vavraia culicis	AJ252961	1364	1364	0
Visvesvaria acridophagus	AF024658	1399	1399	0
Weiseria palustris	AF132544	1402	1374	28
Xenopus laevis	X04025	1826	1826	0

23S Ribosomal RNA (23S rRNA)	**Accessions**	**Length**	**AUGC**	**Other**
Archaea				
Archaeoglobus fulgidus	M64487	2933	2932	1
Desulfurococcus mobilis	X05480	3699	3699	0
Haloarcula marismortui	X13738	2925	2925	0
Haloarcula marismortui rrnA	AF034619	2930	2930	0
Haloarcula marismortui rrnB	AF034620	2930	2930	0
Halobacterium salinarum	X03407	2905	2905	0
Halococcus morrhuae	X05481	2927	2927	0
Haloferax volcanii	**NPA**	2912	2905	7
Methanococcus jannaschii	U67517	3012	3012	0
Methanococcus vannielii	X02729	2958	2958	0
Methanospirillum hungatei	M81323	2910	2909	1
Methanothermobacter thermautotrophicus	X05482	3019	3019	0
Sulfolobus acidocaldarius	M67495	3044	3044	0
Thermococcus celer	M67497	3029	3029	0
Thermofilum pendens	X14835	3069	3069	0
Thermoplasma acidophilum	M32298	2908	2908	0
Thermoproteus tenax	**NPA**	3031	3031	0
Bacteria	**Accessions**	**Length**	**AUGC**	**Other**
Acinetobacter calcoaceticus	X87280	2903	2903	0
Aeromonas hydrophila	X87281	2895	2894	1
Aquifex aeolicus	AE000709	2956	2956	0
Bacillus anthracis	X64645	2920	2887	33
Bacillus sp.	X60981	2914	2914	0
Bacillus subtilis	K00637	2927	2927	0
Bartonella bacilliformis	L39095	2821	2821	0
Bordetella bronchiseptica	X70371	2880	2879	1
Bordetella pertussis	X68323	3009	3007	2
Borrelia burgdorferi	M88330	2926	2926	0
Bradyrhizobium japonicum	Z35330	2874	2874	0
Bradyrhizobium japonicum	X71840	2882	2882	0
Burkholderia cepacia	X16368	2878	2878	0
Burkholderia mallei	Y17183	2882	2882	0

Burkholderia pseudomallei	Y17184	2882	2882	0
Campylobacter coli	U09611	3057	3057	0
Campylobacter jejuni	Z29326	2907	2907	0
Chlamydia suis	U68420	2940	2940	0
Chlamydia trachomatis	U68443	2941	2941	0
Chlamydophila pneumoniae	U76711	2940	2940	0
Chlamydophila psittaci	U68447	2942	2942	0
Chlorobium limicola	M62805	2897	2717	180
Citrobacter freundii	U77928	2906	2896	10
Clostridium botulinum A	X65602	2897	2896	1
Clostridium botulinum B	M94178	2909	2908	1
Clostridium botulinum B	M94259	2909	2907	2
Clostridium botulinum E	M94261	2910	2909	1
Coxiella burnetii	X79704	3204	3204	0
Deinococcus radiodurans	AE002087	2880	2879	1
Deinococcus radiodurans	AE001886	2880	2879	1
Enterococcus faecalis	X79341	2913	2909	4
Erysipelothrix rhusiopathiae	AB019250	2901	2901	0
Escherichia coli	J01695	2904	2904	0
Flexibacter flexilis	M62806	2872	2788	84
Frankia sp.	M55343	3099	3099	0
Geobacillus stearothermophilus	K02663	2928	2928	0
Haemophilus influenzae (operons A-F)	U32742	2897	2897	0
Helicobacter pylori	U27270	2968	2968	0
Heliobacterium chlorum	**NPA**	3001	2986	15
Klebsiella pneumoniae	X87284	2903	2903	0
Lactobacillus delbrueckii	X68426	2909	2909	0
Lactococcus lactis	X68434	2901	2901	0
Leptospira interrogans	X14249	2958	2958	0
Listeria monocytogenes	X64533	2928	2928	0
Listeria monocytogenes	X68420	2932	2932	0
Micrococcus luteus	X06484	3094	3094	0
Mycobacterium leprae	X56657	3122	3122	0
Mycobacterium tuberculosis	Z73902	3138	3138	0
Mycoplasma genitalium	U39694	2917	2917	0
Mycoplasma pneumoniae	X68422	2905	2905	0
Myroides odoratus	M62807	2870	2668	202
Neisseria gonorrhoeae	X67293	2890	2890	0
Neisseria meningitidis	X67300	2890	2890	0
Parachlamydia acanthamoebae	Y07555	3029	3029	0
Pirellula marina	X07408	2885	2885	0
Plesiomonas shigelloides	X65487	2893	2761	132
Pseudomonas aeruginosa	Y00432	2893	2893	0
Rhodobacter capsulatus	X06485	2884	2884	0
Rhodobacter sphaeroides	X53853	2883	2883	0
Rhodococcus erythropolis	AF001265	3133	3125	8
Rhodopseudomonas palustris	X71839	2826	2826	0
Rickettsia prowazekii	AJ235270	2763	2763	0

Rickettsia rickettsii	U11022	2759	2695	64
Ruminobacter amylophilus	X06765	2867	2867	0
Simkania negevensis	U68460	2943	2943	0
Staphylococcus aureus	X68425	2923	2923	0
Staphylococcus carnosus	X68419	2924	2924	0
Streptomyces ambofaciens	M27245	3120	3120	0
Streptomyces griseus	X61478	3120	3120	0
Synechococcus sp. PCC 6301	X00512	2876	2876	0
Thermotoga maritima	M67498	3023	3023	0
Thermus aquaticus	**NPA**	2915	2915	0
Thermus thermophilus	X12612	2915	2915	0
Treponema pallidum (rRNA A)	AE001204	2953	2953	0
Tropheryma whipplei	AF190687	3098	3098	0

Eukaryotic Chloroplast	**Accessions**	**Length**	**AUGC**	**Other**
Alnus incana	M75719	2914	2914	0
Astasia longa	X14386	3105	3105	0
Chlamydomonas eugametos	Z17234	3204	3204	0
Chlamydomonas frankii	L43352	2892	2892	0
Chlamydomonas geitleri	L43353	2921	2921	0
Chlamydomonas gelatinosa	Z15151	2904	2904	0
Chlamydomonas humicola	L42989	2910	2910	0
Chlamydomonas indica	X68893	3127	3127	0
Chlamydomonas iyengarii	L43354	2966	2966	0
Chlamydomonas komma	L43502	2899	2899	0
Chlamydomonas mexicana	L49148	2919	2919	0
Chlamydomonas pallidostigmatica	L43503	3182	3182	0
Chlamydomonas peterfii	L43538	2960	2960	0
Chlamydomonas pitschmannii	Z15152	3017	3017	0
Chlamydomonas reinhardtii	X15727	2902	2902	0
Chlamydomonas sp. SAG 66.72	L43539	3133	3133	0
Chlamydomonas starrii	L43504	3132	3132	0
Chlamydomonas zebra	L43356	2955	2955	0
Chlorella ellipsoidea	M36158	3207	3207	0
Conopholis americana	X59768	2912	2912	0
Epifagus virginiana	X62099	2908	2908	0
Euglena gracilis	X12890	2877	2877	0
Marchantia polymorpha	X04465	2914	2914	0
Nicotiana tabacum	Z00044	2913	2913	0
Odontella sinensis	Z67753	2906	2906	0
Oryza sativa	X15901	2979	2979	0
Palmaria palmata	Z18289	2885	2885	0
Pisum sativum	X55033	2917	2917	0
Plasmodium falciparum (plastid-like)	X61660	2700	2700	0
Toxoplasma gondii (plastid-like)	U18086	2899	2899	0
Zea mays	Z00028	2981	2981	0

Eukaryotic Mitochondrion	**Accessions**	**Length**	**AUGC**	**Other**

Acanthamoeba castellanii	U03732	2719	2719	0
Aedes albopictus	X01078	1335	1335	0
Albinaria caerulea	X83390	1035	1035	0
Albinaria turrita	X71393	1077	1077	0
Allomyces macrogynus	U41288	3162	3162	0
Antilocapra americana	M55540	1573	1573	0
Apis mellifera	L06178	1372	1372	0
Artemia salina	X12965	1148	1137	11
Ascaris suum	X54253	960	960	0
Aspergillus nidulans	J01390	3112	3112	0
Balaenoptera musculus	X72204	1575	1575	0
Bos taurus	J01394	1571	1571	0
Cacozeliana lacertina	AF101007	1341	1341	0
Caenorhabditis elegans	X54252	953	953	0
Cafeteria roenbergensis	AF193903	2587	2587	0
Capra hircus	M55541	1572	1572	0
Cepaea nemoralis	U23045	1215	1215	0
Cervus unicolor	M35875	1572	1572	0
Chlamydomonas eugametos	AF008237	1915	1915	0
Chlamydomonas reinhardtii	X54860	2419	2419	0
Chondrus crispus	Z46224	2583	2583	0
Crithidia fasciculata	X02548	1141	1141	0
Crithidia oncopelti	X51736	1124	1124	0
Crossostoma lacustre	M91245	1680	1680	0
Damaliscus pygargus	M86499	1586	1586	0
Dictyostelium discoideum	D16466	2872	2872	0
Didelphis virginiana	Z29573	1570	1570	0
Drosophila melanogaster	X53506	1335	1324	11
Drosophila yakuba	X03240	1326	1326	0
Equus caballus	X79547	1581	1581	0
Euhadra herklotsi	Z71693	1024	1022	2
Gallus gallus	X52392	1621	1621	0
Homo sapiens	D38112	1558	1558	0
Hydropotes inermis	M35876	1567	1567	0
Katharina tunicata	U09810	1275	1275	0
Leishmania tarentolae	X02354	1173	1173	0
Leptomonas sp.	J03814	1162	1162	0
Locusta migratoria	X80245	1314	1314	0
Loligo bleekeri	AB009838	1333	1302	31
Lumbricus terrestris	U24570	1279	1245	34
Marchantia polymorpha	M68929	2799	2799	0
Meloidogyne javanica	L76261	1512	1512	0
Muntiacus reevesi	M35877	1568	1568	0
Mus musculus	J01420	1582	1581	1
Mytilus edulis	M83756	1244	1242	2
Neurospora crassa	X55443	3465	3465	0
Ochromonas danica	AF287134	2591	2591	0
Odocoileus virginianus	M35874	1567	1567	0

Oenothera berteriana	X02559	3242	3242	0
Pan troglodytes	D38113	1558	1558	0
Paracentrotus lividus	J04815	1551	1551	0
Paracrostoma paludiformis	AF101008	1360	1359	1
Paramecium primaurelia	K00634	2634	2633	1
Paramecium tetraurelia	K01749	2568	2568	0
Pecten maximus	X92688	1411	1411	0
Penicillium chrysogenum	D13859	3416	3416	0
Phoca vitulina	X63726	1565	1565	0
Physarum polycephalum	AF080601	2779	2779	0
Pichia canadensis	D31785	3085	3085	0
Podospora anserina	X14735	3715	3715	0
Porphyra purpurea	AF114794	2590	2590	0
Prototheca wickerhamii	X68722	3004	3004	0
Pylaiella littoralis	Z48620	2693	2693	0
Pyura stolonifera	X74513	1556	1556	0
Rana catesbeiana	X12841	1585	1585	0
Rattus norvegicus	J01438	1559	1559	0
Reclinomonas americana	AF007261	2751	2751	0
Rhizopus stolonifer	NPA	2876	2876	0
Rhodomonas salina	AF288090	2663	2663	0
Saccharomyces cerevisiae	J01527	3273	3273	0
Sceloporus undulatus	L28075	1523	1506	17
Schizosaccharomyces pombe	X06597	2826	2826	0
Strongylocentrotus purpuratus	X12631	1525	1525	0
Suillus sinuspaulianus	L47585	4216	4216	0
Tetrahymena pyriformis	M58010	2595	2595	0
Tetrahymena pyriformis	M58011	2595	2595	0
Tragulus napu	M55539	1575	1575	0
Triticum aestivum	Z11889	3467	3467	0
Trypanosoma brucei	X02547	1152	1152	0
Xenopus laevis	M10217	1640	1640	0
Zea mays	K01868	3514	3514	0

Eukaryotic Nuclear	**Accessions**	**Length**	**AUGC**	**Other**
Aedes albopictus	L22060	4262	4262	0
Arabidopsis thaliana	X52320	3539	3539	0
Babesia bigemina	NPA	3433	3433	0
Brassica napus	D10840	3540	3540	0
Caenorhabditis elegans	X03680	3662	3662	0
Candida albicans	L28817	3513	3513	0
Chlorella ellipsoidea	D17810	3512	3512	0
Crithidia fasciculata	Y00055	4077	4077	0
Dictyostelium discoideum	X00601	3762	3403	359
Didymium iridis	X60210	3857	3857	0
Drosophila melanogaster	M21017	4123	4123	0
Encephalitozoon cuniculi	AJ005581	2552	2506	46
Entamoeba histolytica	X65163	3674	3615	59

Euglena gracilis	X53361	4052	4052	0
Filobasidiella neoformans var. bacillispora	L14067	3544	3544	0
Fragaria x ananassa	X15589	3541	3541	0
Giardia ardeae	X58290	2826	2826	0
Giardia intestinalis	X52949	2850	2848	2
Giardia muris	X65063	2811	2811	0
Herdmania momus	X53538	3721	3721	0
Hexamita inflata	**NPA**	3100	3100	0
Homo sapiens	J01866	5184	5184	0
Lycopersicon esculentum	X13557	3540	3540	0
Microsporidium 57864	U90885	2567	2544	23
Mucor racemosus	M26190	3629	3471	158
Mus musculus	J01871	4869	4869	0
Naegleria gruberi	**NPA**	3615	3615	0
Nosema apis	U76706	2555	2555	0
Nosema apis	U97150	2555	2481	74
Oryza sativa	M11585	3541	3541	0
Physarum polycephalum	V01159	3943	3943	0
Phytophthora megasperma	X75631	3860	3860	0
Plasmodium falciparum (A gene)	U21939	3946	3946	0
Plasmodium falciparum (S gene)	U48228	4381	4381	0
Plasmodium vivax (A gene)	**NPA**	4058	4052	6
Plasmodium vivax (O gene)	**NPA**	4427	4424	3
Plasmodium vivax (S gene)	**NPA**	5461	5460	1
Pneumocystis carinii	M86760	3503	3503	0
Prorocentrum micans	M14649	3562	3562	0
Rattus norvegicus	J01881	4943	4943	0
Saccharomyces cerevisiae	U53879	3554	3554	0
Schizosaccharomyces japonicus	Z32848	3578	3578	0
Schizosaccharomyces pombe	J01359	3661	3656	5
Sinapis alba	X15915	3554	3554	0
Tetrahymena thermophila	X54512	3501	3501	0
Theileria parva	L28036	3523	3396	127
Toxoplasma gondii (P)	X75453	3627	3627	0
Trepomonas agilis	AF015455	3038	2690	348
Trypanosoma brucei	X05682	4188	4188	0
Vairimorpha necatrix	**NPA**	2585	2511	74
Xenopus laevis	X59734	4273	4273	0
Zea mays	**NPA**	3530	3530	0

Appendix B

B.1 Sequence Identity Data Tables for the 16S and 23S Ribosomal RNA Sequences in the Comparatively Predicted Structure Database

Web Reference:

http://www.rna.ccbb.utexas.edu/SIM/4C/mfold_Eval/seq_biases

	Archaea		Bacteria		Eukaryotic Nuclear		Eukaryotic Chloroplast		Eukaryotic Mitochondrion	
16S rRNA										
Sequences	*23*		*195*		*133*		*33*		*112*	
100%	2	0%	24	0%	0	0%	0	0%	0	0%
95%	14	3%	1,564	4%	212	1%	4	0%	12	0%
90%	10	2%	748	2%	138	1%	18	2%	22	0%
85%	46	9%	1,452	4%	312	2%	80	8%	26	0%
80%	54	11%	1,586	4%	320	2%	208	20%	40	0%
75%	166	33%	7,750	20%	682	4%	228	22%	146	1%
70%	154	30%	18,022	48%	2,164	12%	154	15%	304	2%
65%	60	12%	5,650	15%	2,308	13%	58	5%	326	3%
60%	0	0%	844	2%	1,100	6%	56	5%	1,088	9%
55%	0	0%	186	0%	990	6%	86	8%	408	3%
50%	0	0%	4	0%	830	5%	66	6%	208	2%
45%	0	0%	0	0%	872	5%	82	8%	300	2%
40%	0	0%	0	0%	1,686	10%	14	1%	844	7%
35%	0	0%	0	0%	4,916	28%	2	0%	1,410	11%
30%	0	0%	0	0%	782	4%	0	0%	1,192	10%
<30%	0	0%	0	0%	244	1%	0	0%	6,106	49%
Total Pairs	*504*		*37,830*		*17,556*		*1,056*		*12,432*	
23S rRNA										
Sequences	*17*		*75*		*52*		*31*		*81*	
100%	0	0%	0	0%	0	0%	0	0%	0	0%
95%	6	2%	42	1%	8	0%	18	2%	8	0%
90%	0	0%	28	1%	18	1%	58	6%	12	0%
85%	0	0%	90	2%	34	1%	84	9%	40	1%
80%	30	11%	176	3%	24	1%	60	6%	26	0%
75%	10	4%	284	5%	16	1%	158	17%	36	1%
70%	30	11%	434	8%	40	2%	256	28%	134	2%
65%	64	24%	2,864	52%	110	4%	122	13%	64	1%
60%	114	42%	1,476	27%	164	6%	46	5%	56	1%
55%	18	7%	154	3%	106	4%	14	2%	160	2%
50%	0	0%	2	0%	172	6%	52	6%	114	2%
45%	0	0%	0	0%	494	19%	50	5%	198	3%
40%	0	0%	0	0%	618	23%	12	1%	282	4%
35%	0	0%	0	0%	546	21%	0	0%	690	11%
30%	0	0%	0	0%	238	9%	0	0%	652	10%
<30%	0	0%	0	0%	64	2%	0	0%	4,008	62%
Total Pairs	*272*		*5,550*		*2,652*		*930*		*6,480*	

Appendix C

C.1 MFOLD PREDICTION ACCURACY FOR ALL 1,411 RNA SEQUENCES IN THE COMPARATIVELY PREDICTED STRUCTURE DATABASE

Columns:

Accession: Genbank Identifier

Comp BP: Total comparative base pairs (canonical only, G:C, A:U, G:U)

PC: Total comparative base pairs predicted correctly in the Mfold minimum free energy prediction

Acc: Accuracy as a percentage of the total comparative base pairs

Accession: Genbank Identifier

NPA: Not Publically Available (Accession Column)

Web Reference:

http://www.rna.ccbb.utexas.edu/SIM/4C/mfold_Eval/accuracy_efn2

Transfer RNA (tRNA)	Accession	Comp BP	PC	Acc
Archaea				
Archaeoglobus fulgidus (Ala:A)	AE000965	21	7	33%
Halobacterium salinarum (His:H)	X03198	21	18	86%
Halobacterium salinarum (Asn:N)	X03195	22	16	73%
Halobacterium salinarum (Gln:Q)	X03196	21	18	86%
Halobacterium salinarum (Val:V)	K02505	20	12	60%
Halobacterium salinarum (Val:V)	K00244	20	12	60%
Halobacterium sp. NRC-1 (Ala:A)	AE005128	21	12	57%
Halobacterium sp. NRC-1 (Cys:C)	AE005128	20	18	90%
Halobacterium sp. NRC-1 (Gly:G)	AE005077	21	18	86%
Halobacterium sp. NRC-1 (Arg:R)	AE004980	21	18	86%
Haloferax volcanii (Ala:A)	K02507	21	11	52%
Haloferax volcanii (Ala:A)	K02506	21	15	71%
Haloferax volcanii (Ala:A)	K02508	21	20	95%
Haloferax volcanii (Cys:C)	X02584	19	17	89%
Haloferax volcanii (Asp:D)	K00170	21	11	52%
Haloferax volcanii (Glu:E)	K00190	21	13	62%
Haloferax volcanii (Glu:E)	K02510	21	11	52%

Haloferax volcanii (Phe:F)	K02511	21	12	57%
Haloferax volcanii (Gly:G)	K02515	21	21	100%
Haloferax volcanii (Gly:G)	K02513	21	18	86%
Haloferax volcanii (Gly:G)	K02514	22	21	95%
Haloferax volcanii (His:H)	K02516	21	13	62%
Haloferax volcanii (Ile:I)	K00219	21	9	43%
Haloferax volcanii (Lys:K)	K02518	22	13	59%
Haloferax volcanii (Pro:P)	K02521	21	13	62%
Haloferax volcanii (Pro:P)	K02522	20	12	60%
Haloferax volcanii (Gln:Q)	K00183	21	13	62%
Haloferax volcanii (Arg:R)	K00154	21	13	62%
Haloferax volcanii (Arg:R)	K02524	22	12	55%
Haloferax volcanii (Thr:T)	K02526	21	21	100%
Haloferax volcanii (Thr:T)	K02525	21	21	100%
Haloferax volcanii (Val:V)	K02527	20	20	100%
Haloferax volcanii (Val:V)	K00245	20	17	85%
Haloferax volcanii (Trp:W)	K02528	22	13	59%
Haloferax volcanii (Tyr:Y)	K00268	22	20	91%
Methanobacterium formicicum (Asp:D)	AF443995	21	12	57%
Methanocaldococcus jannaschii (Glu:E)	U67517	21	17	81%
Methanocaldococcus jannaschii (Phe:F)	U67517	22	12	55%
Methanocaldococcus jannaschii (His:H)	U67517	21	17	81%
Methanocaldococcus jannaschii (Ile:I)	U67517	22	17	77%
Methanocaldococcus jannaschii (Pro:P)	U67537	21	12	57%
Methanocaldococcus jannaschii (Gln:Q)	U67528	21	13	62%
Methanocaldococcus jannaschii (Arg:R)	U67492	20	15	75%
Methanocaldococcus jannaschii (Thr:T)	U67528	21	21	100%
Methanocaldococcus jannaschii (Val:V)	U67538	20	16	80%
Methanococcus maripaludis (Lys:K)	AF108356	22	17	77%
Methanococcus vannielii (Ala:A)	X00083	21	21	100%
Methanococcus vannielii (Asp:D)	X00916	21	13	62%
Methanococcus vannielii (Pro:P)	X00916	21	12	57%
Methanococcus vannielii (Thr:T)	X00916	20	20	100%
Methanococcus vannielii (Thr:T)	X00916	21	21	100%
Methanococcus vannielii (Thr:T)	X00916	22	13	59%
Methanococcus vannielii (Tyr:Y)	X00916	21	21	100%
Methanosaeta concilii (Ala:A)	X51423	21	21	100%
Methanospirillum hungatei (Ala:A)	M19342	21	16	76%
Methanothermobacter thermautotrophicus (Ala:A)	AE000940	21	20	95%
Methanothermobacter thermautotrophicus (Gly:G)	X06787	20	17	85%
Methanothermobacter thermautotrophicus (Asn:N)	X06788	22	12	55%
Methanothermus fervidus (Ala:A)	M32222	21	12	57%
Methanothermus fervidus (Asp:D)	M26977	21	17	81%
Methanothermus fervidus (Glu:E)	M26978	21	13	62%
Methanothermus fervidus (Ile:I)	M26978	22	17	77%
Methanothermus fervidus (Lys:K)	M26977	22	22	100%
Methanothermus fervidus (Asn:N)	M26978	22	18	82%
Methanothermus fervidus (Pro:P)	M26977	21	11	52%

Methanothermus fervidus (Thr:T)	M26977	22	18	82%
Pyrobaculum aerophilum (Ala:A)	AE009773	21	7	33%
Pyrobaculum aerophilum (Ala:A)	AE009773	21	21	100%
Sulfolobus solfataricus (Phe:F)	AE006696	22	13	59%
Sulfolobus solfataricus (Val:V)	X06054	21	12	57%
Thermococcus sp. MZ12 (Ala:A)	AY017180	21	12	57%
Thermofilum pendens (Gly:G)	X14835	21	21	100%
Thermofilum pendens (Met:M)	X14835	22	7	32%
Thermofilum pendens (Met:M)	X14835	20	15	75%
Thermofilum pendens (Met:M)	X14835	20	15	75%
Thermoplasma acidophilum (Met:M)	K00302	21	7	33%

Bacteria	**Accession**	**Comp BP**	**PC**	**Acc**
Acholeplasma laidlawii (Trp:W)	X15508	21	11	52%
Acidithiobacillus ferrooxidans (Ala:A)	X07395	21	21	100%
Aeromonas hydrophila (His:H)	X12977	21	21	100%
Aeromonas hydrophila (Pro:P)	X12977	21	20	95%
Aeromonas hydrophila (Arg:R)	X12977	21	21	100%
Agrobacterium tumefaciens str. C58 (Ala:A)	AE009341	21	21	100%
Bacillus halodurans (Trp:W)	AP001510	21	15	71%
Bacillus megaterium (Glu:E)	AF142677	21	21	100%
Bacillus megaterium (Lys:K)	AF142677	21	15	71%
Bacillus megaterium (Lys:K)	AF142677	21	15	71%
Bacillus megaterium (Lys:K)	AF142677	21	20	95%
Bacillus megaterium (Arg:R)	AF142677	21	7	33%
Bacillus sporothermodurans (Ala:A)	AF071855	21	20	95%
Bacillus subtilis (Ala:A)	K00141	21	20	95%
Bacillus subtilis (Cys:C)	Z99108	21	11	52%
Bacillus subtilis (Phe:F)	K00637	21	21	100%
Bacillus subtilis (Gly:G)	K00637	21	21	100%
Bacillus subtilis (Gly:G)	K00637	21	17	81%
Bacillus subtilis (Gly:G)	Z99108	21	17	81%
Bacillus subtilis (His:H)	Z99108	20	12	60%
Bacillus subtilis (His:H)	K00637	20	12	60%
Bacillus subtilis (Ile:I)	K00637	21	11	52%
Bacillus subtilis (Ile:I)	Z99104	22	7	32%
Bacillus subtilis (Ile:I)	Z99104	21	12	57%
Bacillus subtilis (Ile:I)	K00637	21	7	33%
Bacillus subtilis (Ile:I)	Z99104	21	21	100%
Bacillus subtilis (Met:M)	K00637	21	11	52%
Bacillus subtilis (Met:M)	K00637	21	16	76%
Bacillus subtilis (Met:M)	K00297	21	7	33%
Bacillus subtilis (Asn:N)	K00637	21	20	95%
Bacillus subtilis (Pro:P)	K00637	21	21	100%
Bacillus subtilis (Gln:Q)	Z99108	20	17	85%
Bacillus subtilis (Arg:R)	K00156	21	21	100%
Bacillus subtilis (Thr:T)	Z99104	21	21	100%

Bacillus subtilis (Thr:T)	K00637	21	15	71%
Bacillus subtilis (Val:V)	K00637	21	12	57%
Campylobacter coli (Ala:A)	AF146727	20	9	45%
Caulobacter crescentus (Ala:A)	L00194	21	13	62%
Cyanophora paradoxa (Ala:A)	M19493	21	21	100%
Cyanophora paradoxa (Glu:E)	U30821	19	18	95%
Cyanophora paradoxa (Gly:G)	X51421	21	21	100%
Cyanophora paradoxa (Ile:I)	M19493	21	21	100%
Cyanophora paradoxa (Ile:I)	M19493	21	21	100%
Cyanophora paradoxa (Ile:I)	M19493	21	11	52%
Cyanophora paradoxa (Ile:I)	M19493	21	20	95%
Cyanophora paradoxa (Ile:I)	M19493	21	20	95%
Escherichia coli (Ala:A)	K00139	20	19	95%
Escherichia coli (Asp:D)	AJ316554	21	6	29%
Escherichia coli (Glu:E)	X05359	22	21	95%
Escherichia coli (Glu:E)	K00188	22	21	95%
Escherichia coli (Phe:F)	AF461394	21	21	100%
Escherichia coli (Met:M)	K00296	20	10	50%
Escherichia coli (Pro:P)	U00039	21	20	95%
Escherichia coli (Arg:R)	K00152	21	12	57%
Escherichia coli (Thr:T)	V00334	21	21	100%
Escherichia coli K12 (Val:V)	AE000262	21	7	33%
Geobacillus stearothermophilus (Phe:F)	K00332	21	21	100%
Geobacillus stearothermophilus (Val:V)	K01065	21	12	57%
Haemophilus influenzae Rd (Gly:G)	U32698	21	15	71%
Lactobacillus delbrueckii (Asp:D)	X15246	20	16	80%
Lactobacillus delbrueckii (Glu:E)	X15246	19	16	84%
Lactobacillus delbrueckii (Asn:N)	X15245	21	20	95%
Lactobacillus delbrueckii (Pro:P)	X15245	21	21	100%
Lactobacillus delbrueckii (Arg:R)	X15246	20	12	60%
Lactobacillus delbrueckii (Val:V)	X15246	21	12	57%
Lactobacillus sakei (Gly:G)	AF401668	21	16	76%
Mycobacterium tuberculosis CDC1551 (Val:V)	AE007103	21	20	95%
Mycoplasma capricolum (Cys:C)	X16746	19	14	74%
Mycoplasma capricolum (Asp:D)	X16745	21	21	100%
Mycoplasma capricolum (Glu:E)	X16748	20	7	35%
Mycoplasma capricolum (Gly:G)	X16749	20	0	0%
Mycoplasma capricolum (His:H)	X16750	20	4	20%
Mycoplasma capricolum (Lys:K)	X16756	20	11	55%
Mycoplasma capricolum (Met:M)	X16758	21	21	100%
Mycoplasma capricolum (Asn:N)	X16744	22	13	59%
Mycoplasma capricolum (Gln:Q)	X16747	19	12	63%
Mycoplasma capricolum (Thr:T)	X16764	21	11	52%
Mycoplasma capricolum (Thr:T)	X16765	21	15	71%
Mycoplasma capricolum (Thr:T)	X16764	21	20	95%
Mycoplasma capricolum (Val:V)	X16769	21	16	76%
Mycoplasma capricolum (Trp:W)	X16766	21	16	76%
Mycoplasma capricolum (Trp:W)	X16767	21	7	33%

Mycoplasma capricolum (Trp:W)	X16767	21	21	100%
Mycoplasma mycoides (Ala:A)	X03154	21	15	71%
Mycoplasma mycoides (Gly:G)	M21590	21	12	57%
Mycoplasma mycoides (Ile:I)	X03154	21	12	57%
Mycoplasma mycoides (Ile:I)	Y00372	21	8	38%
Mycoplasma mycoides (Met:M)	X03154	21	16	76%
Mycoplasma mycoides (Pro:P)	X03154	21	11	52%
Mycoplasma mycoides (Arg:R)	X03154	20	5	25%
Mycoplasma pneumoniae (Gly:G)	AE000043	22	12	55%
Mycoplasma pneumoniae (Lys:K)	AE000043	21	16	76%
Mycoplasma pneumoniae (Gln:Q)	AE000043	20	12	60%
Mycoplasma sp. (Phe:F)	X01305	21	12	57%
Mycoplasma sp. PG50 (Lys:K)	X05660	21	21	100%
Pectobacterium carotovorum subsp. (Ile:I)	AF448597	21	7	33%
Photobacterium phosphoreum (His:H)	X12976	21	21	100%
Pseudomonas aeruginosa (Gly:G)	AE004843	21	21	100%
Pseudomonas aeruginosa (Thr:T)	AF331071	20	15	75%
Pseudomonas aeruginosa (Thr:T)	AE004843	21	21	100%
Pylaiella littoralis (Ala:A)	X14875	21	11	52%
Pylaiella littoralis (Ile:I)	X14875	21	20	95%
Rhodospirillum rubrum (Phe:F)	K00331	21	21	100%
Salmonella typhimurium (Pro:P)	AE008893	20	12	60%
Salmonella typhimurium LT2 (Ala:A)	AE008786	21	20	95%
Salmonella typhimurium LT2 (Cys:C)	AE008895	20	16	80%
Salmonella typhimurium LT2 (Glu:E)	AE008839	22	13	59%
Salmonella typhimurium LT2 (Gly:G)	AE008904	21	21	100%
Salmonella typhimurium LT2 (Gly:G)	AE008904	21	15	71%
Salmonella typhimurium LT2 (Gly:G)	AE008883	21	15	71%
Salmonella typhimurium LT2 (Gly:G)	AE008809	20	16	80%
Salmonella typhimurium LT2 (His:H)	AE008789	21	21	100%
Salmonella typhimurium LT2 (Lys:K)	AE008799	22	11	50%
Salmonella typhimurium LT2 (Asn:N)	AE008883	21	12	57%
Salmonella typhimurium LT2 (Pro:P)	AE008727	20	20	100%
Salmonella typhimurium LT2 (Pro:P)	AE008727	21	21	100%
Salmonella typhimurium LT2 (Gln:Q)	AE008829	20	10	50%
Salmonella typhimurium LT2 (Gln:Q)	AE008883	20	12	60%
Salmonella typhimurium LT2 (Gln:Q)	AE008710	20	12	60%
Salmonella typhimurium LT2 (Arg:R)	AE008893	21	12	57%
Salmonella typhimurium LT2 (Arg:R)	AE008893	20	7	35%
Salmonella typhimurium LT2 (Thr:T)	AE008893	20	15	75%
Salmonella typhimurium LT2 (Thr:T)	AE008762	21	21	100%
Salmonella typhimurium LT2 (Thr:T)	AE008809	21	15	71%
Salmonella typhimurium LT2 (Thr:T)	AE008881	19	5	26%
Salmonella typhimurium LT2 (Val:V)	AE008762	21	16	76%
Salmonella typhimurium LT2 (Val:V)	AE008809	21	6	29%
Salmonella typhimurium LT2 (Trp:W)	AE008881	21	21	100%
Spiroplasma melliferum (Ala:A)	X03715	21	7	33%
Spiroplasma melliferum (Cys:C)	X03715	19	11	58%

Spiroplasma melliferum (Asp:D)	X03715	21	12	57%
Spiroplasma melliferum (Phe:F)	X03715	21	12	57%
Spiroplasma melliferum (Ile:I)	X03715	22	17	77%
Spiroplasma melliferum (Met:M)	X03715	21	12	57%
Spiroplasma melliferum (Pro:P)	X03715	21	17	81%
Spiroplasma melliferum (Arg:R)	X03715	21	7	33%
Staphylococcus epidermidis (Gly:G)	K00199	20	17	85%
Staphylococcus epidermidis (Gly:G)	K00200	21	21	100%
Staphylococcus epidermidis (Asn:N)	AF269878	21	20	95%
Staphylococcus epidermidis (Asn:N)	AF269878	22	11	50%
Streptococcus agalactiae (Ala:A)	AF291419	21	20	95%
Streptomyces coelicolor (Lys:K)	AL596030	21	7	33%
Streptomyces coelicolor A3(2) (Cys:C)	AL157953	20	17	85%
Streptomyces coelicolor A3(2) (Gly:G)	AL157953	21	21	100%
Streptomyces coelicolor A3(2) (Asn:N)	AL163003	21	21	100%
Streptomyces coelicolor A3(2) (Asn:N)	AL163003	21	21	100%
Streptomyces coelicolor A3(2) (Val:V)	AL157953	21	20	95%
Synechococcus sp. PCC 7002 (Phe:F)	K02680	21	20	95%
Synechocystis sp. (Glu:E)	M19535	21	21	100%
Thermus thermophilus (Gly:G)	X51824	21	11	52%
Thermus thermophilus (Ile:I)	M25628	21	21	100%
Thermus thermophilus (Thr:T)	X51824	21	16	76%
Thermus thermophilus (Thr:T)	X51824	21	21	100%
Tolypothrix distorta (Ala:A)	AY007689	20	12	60%
Vibrio cholerae (Asn:N)	AE004132	21	16	76%
Vibrio cholerae (Pro:P)	AE004107			

Eukaryotic Chloroplast	**Accession**	**Comp BP**	**PC**	**Acc**
Chlamydomonas moewusii (Thr:T)	X51398	22	13	59%
Chlamydomonas reinhardtii (Ala:A)	J01395	21	21	100%
Chlamydomonas reinhardtii (Cys:C)	X54407	19	14	74%
Chlamydomonas reinhardtii (Glu:E)	X54408	20	18	90%
Chlamydomonas reinhardtii (Glu:E)	L26266	20	18	90%
Chlamydomonas reinhardtii (Gly:G)	J01399	20	14	70%
Chlamydomonas reinhardtii (Trp:W)	X62566	21	20	95%
Chlorella ellipsoidea (Ala:A)	X05693	21	21	100%
Chlorella ellipsoidea (Arg:R)	X15090	20	5	25%
Chlorella pyrenoidosa (Ile:I)	X03848	21	20	95%
Codium fragile (Gly:G)	M26736	21	20	95%
Codium fragile (Met:M)	M26737	21	21	100%
Codium fragile (Arg:R)	M26738	22	13	59%
Cyanidium caldarium (Lys:K)	D17791	21	12	57%
Euglena gracilis (Ala:A)	X70810	21	21	100%
Euglena gracilis (Cys:C)	X70810	21	19	90%
Euglena gracilis (Asp:D)	X70810	21	7	33%
Euglena gracilis (Asp:D)	K00173	21	21	100%
Euglena gracilis (Phe:F)	X70810	21	21	100%

Euglena gracilis (Phe:F)	K00340	21	21	100%
Euglena gracilis (Phe:F)	K00341	20	7	35%
Euglena gracilis (Gly:G)	X70810	21	21	100%
Euglena gracilis (Gly:G)	X70810	20	16	80%
Euglena gracilis (His:H)	X70810	20	20	100%
Euglena gracilis (Ile:I)	X70810	21	20	95%
Euglena gracilis (Lys:K)	X70810	20	20	100%
Euglena gracilis (Met:M)	X70810	21	16	76%
Euglena gracilis (Asn:N)	X70810	21	20	95%
Euglena gracilis (Pro:P)	X70810	21	21	100%
Euglena gracilis (Gln:Q)	X70810	20	12	60%
Euglena gracilis (Arg:R)	X70810	19	4	21%
Euglena gracilis (Thr:T)	X70810	22	17	77%
Euglena gracilis (Val:V)	X70810	21	21	100%
Euglena gracilis (Trp:W)	X70810	21	10	48%
Glycine max (Met:M)	X07377	21	11	52%
Glycine max (Val:V)	X07675	21	12	57%
Guillardia theta (Arg:R)	AF041468	22	20	91%
Lactuca sativa (His:H)	AF426317	20	12	60%
Marchantia polymorpha (Ala:A)	M20942	21	21	100%
Marchantia polymorpha (Asp:D)	X04465	20	7	35%
Marchantia polymorpha (Glu:E)	X04465	21	19	90%
Marchantia polymorpha (Glu:E)	X04465	21	18	86%
Marchantia polymorpha (Phe:F)	X04465	21	21	100%
Marchantia polymorpha (Gly:G)	X01647	21	16	76%
Marchantia polymorpha (Gly:G)	M20952	21	11	52%
Marchantia polymorpha (His:H)	X04465	20	11	55%
Marchantia polymorpha (Ile:I)	X04465	20	12	60%
Marchantia polymorpha (Ile:I)	M20955	21	20	95%
Marchantia polymorpha (Ile:I)	M20955	20	16	80%
Marchantia polymorpha (Lys:k)	M20959	20	12	60%
Marchantia polymorpha (Met:M)	X04465	21	11	52%
Marchantia polymorpha (Asn:N)	X04465	22	21	95%
Marchantia polymorpha (Pro:P)	X04465	18	10	56%
Marchantia polymorpha (Pro:P)	X04465	21	12	57%
Marchantia polymorpha (Gln:Q)	X04465	20	20	100%
Marchantia polymorpha (Arg:R)	X04465	19	5	26%
Marchantia polymorpha (Arg:R)	X04465	21	4	19%
Marchantia polymorpha (Arg:R)	X04465	21	21	100%
Marchantia polymorpha (Thr:T)	X04465	21	21	100%
Marchantia polymorpha (Thr:T)	X04465	22	17	77%
Marchantia polymorpha (Val:V)	X04465	21	21	100%
Marchantia polymorpha (Val:V)	M20972	20	11	55%
Marchantia polymorpha (Trp:W)	X04465	21	16	76%
Medicago sativa (His:H)	AY029748	20	12	60%
Medicago truncatula (Asp:D)	AC093544	21	21	100%
Medicago truncatula (Met:M)	AC093544	21	11	52%
Medicago truncatula (Pro:P)	AC093544	21	12	57%

Medicago truncatula (Arg:R)	AC093544	22	13	59%
Medicago truncatula (Thr:T)	AC093544	21	21	100%
Medicago truncatula (Trp:W)	AC093544	21	16	76%
Mesostigma viride (Lys:K)	AF166114	21	12	57%
Nephroselmis olivacea (Ala:A)	AF137379	18	12	67%
Nicotiana tabacum (Cys:C)	Z00044	20	13	65%
Nicotiana tabacum (Asp:D)	Z00044	21	12	57%
Nicotiana tabacum (Glu:E)	Z00044	21	15	71%
Nicotiana tabacum (Phe:F)	Z00044	21	21	100%
Nicotiana tabacum (Gly:G)	Z00044	20	13	65%
Nicotiana tabacum (His:H)	Z00044	20	12	60%
Nicotiana tabacum (Met:M)	Z00044	21	11	52%
Nicotiana tabacum (Asn:N)	Z00044	22	11	50%
Nicotiana tabacum (Pro:P)	Z00044	21	12	57%
Nicotiana tabacum (Gln:Q)	Z00044	20	10	50%
Nicotiana tabacum (Thr:T)	Z00044	21	21	100%
Nicotiana tabacum (Trp:W)	Z00044	21	15	71%
Nicotiana tabacum (Tyr:Y)	X00360	21	9	43%
Nicotiana tabacum (Tyr:Y)	X00361	21	20	95%
Parodia erinacea (Thr:T)	AY064336	21	16	76%
Pelargonium zonale (Arg:R)	X01120	19	13	68%
Phaseolus vulgaris (Phe:F)	K00336	21	21	100%
Pisum sativum (Phe:F)	X04551	20	18	90%
Pisum sativum (Gly:G)	X05394	22	18	82%
Pisum sativum (Pro:P)	X05395	21	12	57%
Pisum sativum (Arg:R)	M16863	18	5	28%
Pisum sativum (Val:V)	X55033	21	21	100%
Pisum sativum (Trp:W)	X05395	22	13	59%
Pisum sativum (Trp:W)	X05395	21	16	76%
Ptychosperma burretianum (Glu:E)	AF449169	21	15	71%
Scenedesmus obliquus (Phe:F)	M25610	20	11	55%
Scenedesmus obliquus (Met:M)	M25611	20	19	95%
Scenedesmus obliquus (Tyr:Y)	X02224	21	21	100%
Sinapis alba (His:H)	X17331	20	14	70%
Sinapis alba (Gln:Q)	X13558	20	12	60%
Spinacia oleracea (Cys:C)	AJ400848	19	7	37%
Spinacia oleracea (Asp:D)	AJ400848	21	7	33%
Spinacia oleracea (Glu:E)	AJ400848	21	15	71%
Spinacia oleracea (Phe:F)	X02686	21	12	57%
Spinacia oleracea (His:H)	AJ400848	20	20	100%
Spinacia oleracea (Ile:I)	K01839	20	11	55%
Spinacia oleracea (Ile:I)	K00222	21	20	95%
Spinacia oleracea (Ile:I)	K00222	21	20	95%
Spinacia oleracea (Ile:I)	K00222	21	20	95%
Spinacia oleracea (Ile:I)	K02848	20	11	55%
Spinacia oleracea (Met:M)	AJ400848	21	11	52%
Spinacia oleracea (Pro:P)	K00358	21	17	81%
Spinacia oleracea (Pro:P)	AJ400848	21	12	57%

Spinacia oleracea (Arg:R)	AJ400848	21	11	52%
Spinacia oleracea (Thr:T)	AJ400848	21	21	100%
Spinacia oleracea (Thr:T)	K00281	21	21	100%
Spinacia oleracea (Thr:T)	AJ400848	21	12	57%
Spinacia oleracea (Val:V)	AJ400848	21	12	57%
Spinacia oleracea (Val:V)	K00247	20	11	55%
Spinacia oleracea (Val:V)	K00247	20	11	55%
Spinacia oleracea (Trp:W)	K00262	21	15	71%
Spirodela punctata (Arg:R)	X00764	22	11	50%
Triticum aestivum (Met:M)	X02560	19	16	84%
Triticum aestivum (Trp:W)	K02003	20	15	75%
Triticum aestivum (Gly:G)	X00756	20	8	40%
Vicia faba (Glu:E)	X00682	21	15	71%
Vicia faba (Phe:F)	X51471	21	21	100%
Zea mays (Cys:C)	X86563	21	11	52%
Zea mays (Trp:W)	X86563	21	20	95%

Eukaryotic Nuclear	**Accession**	**Comp BP**	**PC**	**Acc**
Arabidopsis thaliana (Asp:D)	AC016041	21	17	81%
Arabidopsis thaliana (Phe:F)	AC011665	20	9	45%
Arabidopsis thaliana (Lys:K)	AC026234	21	8	38%
Arabidopsis thaliana (Pro:P)	NM_105549	21	17	81%
Arabidopsis thaliana (Pro:P)	NM_105549	21	17	81%
Arabidopsis thaliana (Pro:P)	AC018907	21	17	81%
Arabidopsis thaliana (Arg:R)	AB019236	20	9	45%
Arabidopsis thaliana (Val:V)	AC025417	20	12	60%
Bombyx mori (Ala:A)	M23363	21	12	57%
Bombyx mori (Glu:E)	X03602	21	7	33%
Bombyx mori (Gly:G)	K00206	22	12	55%
Bos taurus (Asp:D)	K00175	18	10	56%
Bos taurus (Phe:F)	K00352	21	12	57%
Bos taurus (Arg:R)	V00134	21	7	33%
Bos taurus (Arg:R)	X04541	21	11	52%
Bos taurus (Thr:T)	M26109	20	10	50%
Bos taurus (Trp:W)	M10543	21	21	100%
Bos taurus (Tyr:Y)	M26210	21	21	100%
Caenorhabditis elegans (Asp:D)	U41014	21	7	33%
Caenorhabditis elegans (Lys:K)	AF040661	22	20	91%
Caenorhabditis elegans (Pro:P)	AC024859	21	18	86%
Caenorhabditis elegans (Trp:W)	U70846	21	5	24%
Dictyostelium discoideum (Glu:E)	AF037042	21	20	95%
Dictyostelium discoideum (Val:V)	AF067200	20	11	55%
Dictyostelium discoideum (Val:V)	X03499	19	6	32%
Drosophila melanogaster (Ala:A)	AC009461	21	12	57%
Drosophila melanogaster (Asp:D)	NG_000295	21	12	57%
Drosophila melanogaster (Glu:E)	V00238	21	7	33%
Drosophila melanogaster (Glu:E)	AC010564	21	7	33%

Drosophila melanogaster (Glu:E)	K00193	21	11	52%
Drosophila melanogaster (Glu:E)	NG_000161	21	11	52%
Drosophila melanogaster (Phe:F)	AC023722	21	12	57%
Drosophila melanogaster (Phe:F)	K00349	21	12	57%
Drosophila melanogaster (Gly:G)	NG_000194	22	12	55%
Drosophila melanogaster (Gly:G)	X07778	21	11	52%
Drosophila melanogaster (His:H)	AC099014	21	12	57%
Drosophila melanogaster (His:H)	K00215	21	12	57%
Drosophila melanogaster (Ile:I)	NG_000454	20	16	80%
Drosophila melanogaster (Lys:K)	AC008257	21	16	76%
Drosophila melanogaster (Lys:K)	AC008257	21	16	76%
Drosophila melanogaster (Lys:K)	K01859	21	7	33%
Drosophila melanogaster (Met:M)	K00462	20	7	35%
Drosophila melanogaster (Asn:N)	AC008257	21	11	52%
Drosophila melanogaster (Pro:P)	AC018491	21	16	76%
Drosophila melanogaster (Pro:P)	AE003723	21	16	76%
Drosophila melanogaster (Arg:R)	AC008257	20	12	60%
Drosophila melanogaster (Arg:R)	AC021639	21	16	76%
Drosophila melanogaster (Thr:T)	AC097445	20	20	100%
Drosophila melanogaster (Val:V)	AC009461	20	5	25%
Drosophila melanogaster (Val:V)	AC010713	20	12	60%
Drosophila melanogaster (Val:V)	AC009461	20	17	85%
Drosophila melanogaster (Val:V)	M25880	20	20	100%
Drosophila melanogaster (Val:V)	AC091207	20	11	55%
Drosophila melanogaster (Tyr:Y)	M26124	21	21	100%
Gallus gallus (Lys:K)	J00881	21	16	76%
Homo sapiens (Ala:A)	AC013472	21	20	95%
Homo sapiens (Ala:A)	AL121936	21	20	95%
Homo sapiens (Ala:A)	AL121932	21	16	76%
Homo sapiens (Glu:E)	J00309	20	10	50%
Homo sapiens (Glu:E)	AL355149	20	0	0%
Homo sapiens (Phe:F)	AL662890	21	12	57%
Homo sapiens (Gly:G)	K00208	21	21	100%
Homo sapiens (Gly:G)	K00209	21	18	86%
Homo sapiens (Gly:G)	AL355149	22	12	55%
Homo sapiens (Gly:G)	K00208	22	0	0%
Homo sapiens (His:H)	X01553	21	21	100%
Homo sapiens (His:H)	U43279	21	11	52%
Homo sapiens (His:H)	U43279	20	10	50%
Homo sapiens (His:H)	X01553	21	12	57%
Homo sapiens (Ile:I)	AL121934	20	5	25%
Homo sapiens (Lys:K)	U00939	21	16	76%
Homo sapiens (Asn:N)	AL356957	21	11	52%
Homo sapiens (Asn:N)	K01921	21	12	57%
Homo sapiens (Asn:N)	X15813	21	20	95%
Homo sapiens (Pro:P)	AC024952	21	16	76%
Homo sapiens (Pro:P)	AC008443	21	16	76%
Homo sapiens (Gln:Q)	K01921	21	12	57%

Homo sapiens (Gln:Q)	X15814	20	19	95%
Homo sapiens (Gln:Q)	X15813	21	20	95%
Homo sapiens (Arg:R)	AJ333675	20	7	35%
Homo sapiens (Arg:R)	AL121936	20	12	60%
Homo sapiens (Arg:R)	AC083880	21	7	33%
Homo sapiens (Thr:T)	AL163636	21	20	95%
Homo sapiens (Val:V)	AC008443	20	17	85%
Homo sapiens (Val:V)	AC008443	20	17	85%
Homo sapiens (Val:V)	AL031229	20	17	85%
Homo sapiens (Val:V)	AC008443	20	17	85%
Homo sapiens (Val:V)	AC008443	20	17	85%
Homo sapiens (Val:V)	AC005783	20	12	60%
Homo sapiens (Tyr:Y)	X04779	21	21	100%
Hordeum vulgare (Glu:E)	X06283	22	13	59%
Hordeum vulgare (Glu:E)	X06283	21	7	33%
Hordeum vulgare (Glu:E)	X06378	21	15	71%
Hordeum vulgare (Glu:E)	X06284	21	17	81%
Hordeum vulgare (Gln:Q)	X06376	21	20	95%
Lupinus luteus (Glu:E)	M23387	21	15	71%
Lupinus luteus (Phe:F)	K00345	20	13	65%
Lupinus luteus (Gly:G)	X05493	20	12	60%
Lupinus luteus (His:H)	M16065	21	12	57%
Lupinus luteus (Ile:I)	X06459	20	11	55%
Lupinus luteus (Asn:N)	X07526	21	20	95%
Lupinus luteus (Val:V)	X05082	20	7	35%
Lupinus luteus (Val:V)	X05082	20	11	55%
Mus musculus (Glu:E)	X00229	21	4	19%
Mus musculus (Glu:E)	X00229	21	4	19%
Mus musculus (Gly:G)	AC069308	22	12	55%
Mus musculus (His:H)	J00642	21	12	57%
Mus musculus (Ile:I)	AL589879	20	5	25%
Mus musculus (Met:M)	X04525	20	11	55%
Mus musculus (Asn:N)	AY050218	21	11	52%
Mus musculus (Pro:P)	K00360	21	16	76%
Mus musculus (Gln:Q)	AC092498	21	20	95%
Mus musculus (Gln:Q)	M16252	21	20	95%
Neurospora crassa (Phe:F)	X02710	20	18	90%
Oryctolagus cuniculus (Asp:D)	K00176	21	16	76%
Oryctolagus cuniculus (Lys:K)	K00289	21	11	52%
Oryctolagus cuniculus (Met:M)	X68632	20	11	55%
Oryza sativa (Cys:C)	AC092750	21	11	52%
Oryza sativa (Phe:F)	AC092750	21	21	100%
Oryza sativa (Gly:G)	AC092750	19	16	84%
Oryza sativa (Ile:I)	AC099402	19	18	95%
Oryza sativa (Met:M)	AC092750	21	11	52%
Oryza sativa (Met:M)	AC092750	21	11	52%
Oryza sativa (Asn:N)	AC099402	20	6	30%
Oryza sativa (Arg:R)	AC099402	19	9	47%

Oryza sativa (Thr:T)	AC092750	21	21	100%
Oryza sativa (Thr:T)	AC092750	19	10	53%
Oryza sativa (Val:V)	AC099402	21	12	57%
Pichia jadinii (Ile:I)	K01061	20	11	55%
Pichia jadinii (Pro:P)	K00357	20	12	60%
Pichia jadinii (Tyr:Y)	M24830	20	20	100%
Rattus norvegicus (Asp:D)	K00444	21	7	33%
Rattus norvegicus (Asp:D)	K03129	21	7	33%
Rattus norvegicus (Asp:D)	V01269	21	7	33%
Rattus norvegicus (Glu:E)	V01272	21	4	19%
Rattus norvegicus (Glu:E)	K00446	21	4	19%
Rattus norvegicus (Glu:E)	K00446	20	5	25%
Rattus norvegicus (Glu:E)	K00195	21	12	57%
Rattus norvegicus (Phe:F)	M22764	21	12	57%
Rattus norvegicus (Gly:G)	V01272	21	18	86%
Rattus norvegicus (Gly:G)	X00706	21	18	86%
Rattus norvegicus (Lys:K)	X04545	21	16	76%
Rattus norvegicus (Asn:N)	K00166	21	11	52%
Rattus norvegicus (Pro:P)	K01637	21	5	24%
Rattus norvegicus (Gln:Q)	V01265	21	20	95%
Rattus norvegicus (Val:V)	M34549	20	17	85%
Saccharomyces cerevisiae (Cys:C)	M34549	22	16	73%
Saccharomyces cerevisiae (Cys:C)	X01939	22	16	73%
Saccharomyces cerevisiae (Asp:D)	X90518	21	7	33%
Saccharomyces cerevisiae (Asp:D)	M25168	21	7	33%
Saccharomyces cerevisiae (Glu:E)	U51030	19	15	79%
Saccharomyces cerevisiae (Glu:E)	U18778	20	20	100%
Saccharomyces cerevisiae (Glu:E)	K00191	20	20	100%
Saccharomyces cerevisiae (Phe:F)	M10263	21	20	95%
Saccharomyces cerevisiae (Phe:F)	M14867	21	20	95%
Saccharomyces cerevisiae (Gly:G)	K00204	22	18	82%
Saccharomyces cerevisiae (Gly:G)	U18779	22	18	82%
Saccharomyces cerevisiae (Gly:G)	Z71561	21	18	86%
Saccharomyces cerevisiae (Gly:G)	Z71561	20	16	80%
Saccharomyces cerevisiae (His:H)	M26097	20	10	50%
Saccharomyces cerevisiae (Ile:I)	U18922	20	7	35%
Saccharomyces cerevisiae (Ile:I)	X69098	20	19	95%
Saccharomyces cerevisiae (Lys:K)	K00286	21	4	19%
Saccharomyces cerevisiae (Lys:K)	U18530	21	11	52%
Saccharomyces cerevisiae (Lys:K)	K00287	21	14	67%
Saccharomyces cerevisiae (Met:M)	J01372	21	11	52%
Saccharomyces cerevisiae (Met:M)	M10268	20	20	100%
Saccharomyces cerevisiae (Asn:N)	M26099	20	5	25%
Saccharomyces cerevisiae (Pro:P)	M26096	20	12	60%
Saccharomyces cerevisiae (Gln:Q)	X66375	19	7	37%
Saccharomyces cerevisiae (Gln:Q)	U18796	20	7	35%
Saccharomyces cerevisiae (Arg:R)	U18917	20	4	20%
Saccharomyces cerevisiae (Arg:R)	L47993	21	7	33%

Saccharomyces cerevisiae (Arg:R)	U18530	21	7	33%
Saccharomyces cerevisiae (Arg:R)	K00158	21	7	33%
Saccharomyces cerevisiae (Arg:R)	K00159	21	5	24%
Saccharomyces cerevisiae (Thr:T)	K00279	20	19	95%
Saccharomyces cerevisiae (Val:V)	Z75085	20	12	60%
Saccharomyces cerevisiae (Val:V)	K00249	20	12	60%
Saccharomyces cerevisiae (Val:V)	Z47814	20	10	50%
Saccharomyces cerevisiae (Trp:W)	M35060	22	15	68%
Saccharomyces cerevisiae (Tyr:Y)	M10266	20	20	100%
Saccharomyces pastorianus (Phe:F)	X00655	21	0	0%
Schizosaccharomyces pombe (Asp:D)	AL590457	21	15	71%
Schizosaccharomyces pombe (Glu:E)	AL121794	21	12	57%
Schizosaccharomyces pombe (Phe:F)	Z97208	22	7	32%
Schizosaccharomyces pombe (Phe:F)	K00344	20	7	35%
Schizosaccharomyces pombe (His:H)	AL031825	20	11	55%
Schizosaccharomyces pombe (Lys:K)	Z97185	21	21	100%
Schizosaccharomyces pombe (Arg:R)	X00239	20	7	35%
Schizosaccharomyces pombe (Arg:R)	AL590457	20	12	60%
Schizosaccharomyces pombe (Arg:R)	AL590457	21	21	100%
Schizosaccharomyces pombe (Tyr:Y)	K00273	20	12	60%
Sorghum bicolor (Gly:G)	AF466201	20	12	60%
Tetrahymena pyriformis (Asn:N)	X16643	19	16	84%
Tetrahymena thermophila (Gln:Q)	M35401	21	11	52%
Tetrahymena thermophila (Gln:Q)	M11464	21	11	52%
Tetrahymena thermophila (Gln:Q)	M35400	21	10	48%
Trypanosoma brucei (Lys:K)	AF047724	21	11	52%
Trypanosoma brucei (Val:V)	X16590	17	6	35%
Trypanosoma brucei rhodesiense (Gln:Q)	X16590	15	4	27%
Xenopus laevis (Ala:A)	Y00430	21	17	81%
Xenopus laevis (Asp:D)	X04460	21	7	33%
Xenopus laevis (Phe:F)	K02849	21	7	33%
Xenopus laevis (Lys:K)	Y00163	21	12	57%
Xenopus laevis (Val:V)	X04819	20	14	70%
Xenopus laevis (Val:V)	X04819	20	12	60%

5S Ribosomal RNA (5S rRNA)

Archaea	**Accession**	**Comp BP**	**PC**	**Acc**
Haloarcula marismortui	AF034620	34	29	85%
Haloferax mediterranei	X14441	34	10	29%
Methanocaldococcus jannaschii	U67518	37	28	76%
Methanolobus tindarius	M34910	41	40	98%
Methanothermococcus thermolithotrophicus	M34911	36	18	50%
Methanothermus fervidus	M26976	40	31	78%
Pyrococcus woesei	X15329	39	30	77%
Pyrodictium occultum	M21086	45	41	91%
Sulfolobus acidocaldarius	V01286	41	38	93%

Sulfolobus solfataricus	X01588	41	38	93%
Thermococcus celer	X07692	39	36	92%
Thermoplasma acidophilum	M32297	36	33	92%
Bacteria	**Accession**	**Comp BP**	**PC**	**Acc**
Acidithiobacillus ferrooxidans	M11542	35	32	91%
Agrobacterium tumefaciens	X02627	35	33	94%
Arthrobacter globiformis	M16173	34	25	74%
Arthrobacter globiformis	X08002	33	22	67%
Arthrobacter oxydans	X08000	33	22	67%
Bacillus subtilis	D11460	32	23	72%
Campylobacter jejuni	AL139076	31	9	29%
Deinococcus radiodurans	AE002087	35	31	89%
Delftia acidovorans	AJ131594	34	26	76%
Escherichia coli	V00336	37	10	27%
Geobacillus stearothermophilus	M10816	33	18	55%
Geobacillus stearothermophilus	AJ251080	33	16	48%
Geobacillus stearothermophilus	M24839	29	8	28%
Geobacillus stearothermophilus	M25591	33	27	82%
Haemophilus influenzae	U32688	35	29	83%
Micrococcus luteus	K02682	35	27	77%
Mycoplasma genitalium	U39694	32	26	81%
Planctomyces brasiliensis	M35168	30	6	20%
Pseudomonas aeruginosa	K02353	36	32	89%
Pseudomonas stutzeri	M34776	34	22	65%
Rhodobacter capsulatus	X04585	33	28	85%
Spiroplasma melliferum	X06098	17	6	35%
Sporosarcina pasteurii	X02024	33	6	18%
Staphylococcus aureus	L36472	32	18	56%
Synechococcus sp. PCC 6301	X00757	32	27	84%
Thermus aquaticus	X01590	37	14	38%
Thermus sp.	M16532	34	30	88%
Thermus thermophilus	V01415	36	8	22%
Eukaryotic Chloroplast	**Accession**	**Comp BP**	**PC**	**Acc**
Chlamydomonas reinhardtii	BK000554	34	29	85%
Euglena gracilis	K02483	31	5	16%
Marchantia polymorpha	X00666	34	29	85%
Zea mays	M19943	35	28	80%
Eukaryotic Mitochondrion	**Accession**	**Comp BP**	**PC**	**Acc**
Reclinomonas americana	U59762	34	31	91%
Eukaryotic Nuclear	**Accession**	**Comp BP**	**PC**	**Acc**
Acanthamoeba castellanii	V00003	36	27	75%

Acheta domesticus	M16074	35	29	83%
Amoebidium parasiticum	M36306	36	34	94%
Ascobolus immersus	X99087	35	32	91%
Asterias vulgaris	X00992	32	27	84%
Aurelia aurita	X00991	34	28	82%
Blastocladiella simplex	X01543	36	32	89%
Blepharisma japonicum	J01851	35	31	89%
Bos taurus	X57170	35	29	83%
Branchiostoma belcheri	X13034	35	29	83%
Candida albicans	X00868	35	31	89%
Chlamydomonas reinhardtii	X02706	33	16	48%
Christiansenia pallida	M58383	33	26	79%
Crithidia fasciculata	V00149	34	10	29%
Crypthecodinium cohnii	M25115	36	29	81%
Cryptococcus neoformans var. neoformans	L14753	33	30	91%
Cyanophora paradoxa	M33029	34	24	71%
Diatoma tenue	D00058	35	12	34%
Drosophila melanogaster	M25016	33	26	79%
Dugesia japonica	X01551	35	27	77%
Emplectonema gracile	X00021	34	31	91%
Enchytraeus albidus	X03911	35	29	83%
Equisetum arvense	X00377	33	29	88%
Euglena gracilis	X01484	36	27	75%
Exobasidium vaccinii	X00069	35	32	91%
Globodera pallida	L28955	34	27	79%
Gracilaria compressa	X00999	34	9	26%
Homo sapiens	Z75742	29	8	28%
Hyphodontia paradoxa	X73890	34	27	79%
Mesocricetus auratus	J00063	35	29	83%
Mortierella formosensis	M36312	36	32	89%
Octopus vulgaris	X06835	33	26	79%
Oryza sativa	M18171	33	25	76%
Phaseolus vulgaris	X06843	33	31	94%
Physarum polycephalum	X02036	36	27	75%
Plagiomnium trichomanes	X01619	33	31	94%
Plasmodium falciparum	AF239766	35	10	29%
Pneumocystis carinii	M28193	36	32	89%
Pseudocentrotus depressus	X04307	35	29	83%
Saccharomyces cerevisiae	X67579	35	28	80%
Schizochytrium aggregatum	X06104	36	29	81%
Schizosaccharomyces pombe	K00570	36	33	92%
Spirogyra sp.	M10438	35	0	0%
Tetrahymena thermophila	X00475	36	32	89%
Xenopus laevis	X05089	36	29	81%
16S Ribosomal RNA (16S rRNA)	**Accession**	**Comp BP**	**PC**	**Acc**

Archaea				
Aeropyrum pernix	AP000062	451	292	65%
Archaeoglobus fulgidus	X05567	448	256	57%
Haloarcula marismortui rrnA	X61688	438	254	58%
Haloarcula marismortui rrnB	X61689	440	296	67%
Halobacterium sp.	AE005128	440	225	51%
Haloferax volcanii	K00421	439	336	77%
Methanobacterium formicicum	M36508	438	309	71%
Methanococcus jannaschii	U67517	445	244	55%
Methanococcus vannielii	M36507	437	230	53%
Methanospirillum hungatei	M60880	432	272	63%
Methanothermobacter thermautotrophicus	AE000930	442	275	62%
Natronobacterium innermongoliae	AF009601	442	322	73%
Natronorubrum bangense	Y14028	441	282	64%
Pyrococcus abyssi	AJ248283	453	308	68%
Pyrococcus furiosus	U20163	444	275	62%
Pyrococcus horikoshii	AP000001	452	308	68%
Pyrodictium occultum	M21087	447	258	58%
Sulfolobus acidocaldarius	D14876	446	257	58%
Sulfolobus P2	**NPA**	447	228	51%
Sulfolobus solfataricus	X03235	446	257	58%
Thermococcus celer	M21529	447	330	74%
Thermoplasma acidophilum	AL445067	440	266	60%
Thermoproteus tenax	M35966	456	296	65%

Bacteria	**Accession**	**Comp BP**	**PC**	**Acc**
Acidobacterium capsulatum	D26171	401	176	44%
Acinetobacter calcoaceticus	M34139	434	186	43%
Actinomyces israelii	X82450	407	102	25%
Aeromonas hydrophila	X60407	446	215	48%
Agrobacterium tumefaciens	M11223	429	239	56%
Allochromatium vinosum	M26629	423	178	42%
Anabaena sp.	X59559	432	219	51%
Aquifex aeolicus	AE000709	473	280	59%
Aquifex pyrophilus	M83548	468	282	60%
Arthrobacter globiformis	M23411	446	206	46%
Azorhizobium caulinodans	D11342	424	182	43%
Bacillus anthracis	X55059	393	218	55%
Bacillus cereus	X55060	407	277	68%
Bacillus halodurans	AB013373	450	306	68%
Bacillus subtilis	K00637	451	241	53%
Bacteroides fragilis	M61006	442	290	66%
Bartonella bacilliformis	Z11683	408	193	47%
Bartonella henselae	M73229	410	197	48%
Bartonella quintana	M11927	429	207	48%
Beggiatoa sp. 1401-13	L40997	428	241	56%
Bordetella bronchiseptica	U04948	448	298	67%

Bordetella parapertussis	U04949	421	265	63%
Bordetella pertussis	U04950	420	252	60%
Borrelia burgdorferi	M88329	455	225	49%
Borrelia hermsii	U42292	450	228	51%
Brachyspira hyodysenteriae	U23035	421	196	47%
Bradyrhizobium japonicum	Z35330	432	215	50%
Brevinema andersonii	L31543	408	195	48%
Brucella melitensis	L26166	400	207	52%
Buchnera sp. APS	AP000398	420	237	56%
Burkholderia mallei	S55008	376	170	45%
Burkholderia sp.	U37342	432	241	56%
Campylobacter jejuni	Z29326	442	274	62%
Campylobacter sputorum	X67775	420	209	50%
Chlamydia muridarum	AE002280	458	237	52%
Chlamydia trachomatis	U68443	458	238	52%
Chlamydophila pneumoniae	L06108	460	229	50%
Chlamydophila psittaci	U68447	456	230	50%
Chlorobium vibrioforme	M62791	430	192	45%
Chlorogloeopsis sp. PCC 7518	X68780	432	204	47%
Chromohalobacter marismortui	X87222	431	195	45%
Citrobacter freundii	M59291	416	170	41%
Clostridium botulinum F	L37593	421	266	63%
Clostridium difficile	X73450	423	214	51%
Clostridium innocuum	M23732	440	257	58%
Clostridium perfringens	M69264	438	174	40%
Clostridium tetani	X74770	436	237	54%
Comamonas testosteroni	M11224	441	211	48%
Corynebacterium diphtheriae	X84248	432	155	36%
Coxiella burnetii	M21291	418	208	50%
Cristispira CP1	U42638	430	167	39%
Deferribacter thermophilus	U75602	460	265	58%
Deinococcus radiodurans	M21413	437	273	62%
Desulfovibrio desulfuricans	M34113	447	205	46%
Dichelobacter nodosus	M35016	448	247	55%
Edwardsiella tarda	AF015259	402	247	61%
Enterococcus faecalis	Y18293	407	262	64%
Enterococcus faecium	AF070223	435	271	62%
environ.Eubacteria clone W15	NPA	407	262	64%
Epulopiscium sp.	M99572	412	131	32%
Erysipelothrix rhusiopathiae	M23728	430	259	60%
Escherichia coli	J01695	457	242	53%
Eubacterium brachy	Z36272	425	265	62%
Francisella tularensis	Z21931	439	171	39%
Frankia sp.	M55343	442	200	45%
Fusobacterium necrophorum	X74408	423	174	41%
Fusobacterium nucleatum subsp. nucleatum	M58683	430	284	66%
Gemmata obscuriglobus	X56305	404	223	55%
Geotoga subterranea	L10659	434	203	47%

Gluconacetobacter liquefaciens	X75617	434	160	37%
Haemobartonella felis	U95297	382	135	35%
Haemophilus influenzae	X87977	407	165	41%
Haemophilus influenzae (operons A-F)	U32741	447	199	45%
Halomonas halodenitrificans	L04942	440	148	34%
Helicobacter pylori	M88157	419	195	47%
Heliobacterium chlorum	M11212	436	167	38%
Holophaga foetida	X77215	441	217	49%
Isosphaera pallida	X64372	398	174	44%
Klebsiella pneumoniae	X80684	404	182	45%
Lactobacillus acidophilus	M58802	430	231	54%
Lactococcus lactis subsp. lactis	AE006456	451	275	61%
Legionella pneumophila	M59157	425	191	45%
Leptonema illini	M88719	442	218	49%
Leptospira interrogans	X17547	440	247	56%
Leptospirillum ferriphilum	AF356830	424	199	47%
Listeria monocytogenes	M58822	429	246	57%
Mesorhizobium loti	AP003001	429	199	46%
Methylobacterium sp.	Z23160	409	203	50%
Methylococcus capsulatus	X72771	420	135	32%
Micrococcus luteus	M38242	420	143	34%
Microcystis aeruginosa	U03402	406	239	59%
Mycobacterium avium	X52918	415	144	35%
Mycobacterium leprae	X56657	452	97	21%
Mycobacterium tuberculosis	X52917	418	94	22%
Mycoplasma capricolum	X00921	437	209	48%
Mycoplasma gallisepticum	M22441	434	236	54%
Mycoplasma genitalium	U39694	437	220	50%
Mycoplasma hyopneumoniae	Y00149	443	303	68%
Mycoplasma mycoides	M23943	376	130	35%
Mycoplasma pneumoniae	M29061	420	229	55%
Myxococcus xanthus	M34114	446	224	50%
Neisseria gonorrhoeae	X07714	443	251	57%
Neisseria meningitidis	AE002364	444	264	59%
Nocardia asteroides	X80606	423	156	37%
Oscillatoria agardhii	X84811	420	271	65%
Pasteurella multocida	M35018	435	184	42%
Petrotoga miotherma	L10657	364	213	59%
Pirellula marina	X62912	409	129	32%
Pirellula staleyi	M34126	433	195	45%
Planctomycetaceae Schlesner 670	X81948	427	227	53%
Plesiomonas shigelloides	X74688	411	201	49%
Pleurocapsa sp.	X78681	419	197	47%
Porphyromonas gingivalis	L16492	419	151	36%
Proteus vulgaris	X07652	456	227	50%
Pseudomonas aeruginosa	M34133	424	214	50%
Pseudomonas putida	D84020	448	223	50%
Pseudomonas sp.	U37339	429	231	54%

Psychrobacter pacificensis	AB016054	448	199	44%
Rhodobium orientis	D30792	407	224	55%
Rhodoblastus acidophilus	M34128	414	210	51%
Rhodococcus erythropolis	AF001265	442	186	42%
Rickettsia bellii	U11014	439	244	56%
Rickettsia prowazekii	M21789	439	245	56%
Rickettsia rickettsii	L36217	417	223	53%
Salmonella typhimurium	X80681	447	216	48%
Serratia marcescens	M59160	431	197	46%
Shewanella putrefaciens	X81623	446	206	46%
Shigella dysenteriae	X96966	428	194	45%
Spirochaeta aurantia	M57740	436	251	58%
Staphylococcus aureus	L36472	446	232	52%
Streptobacillus moniliformis	Z35305	428	141	33%
Streptococcus mutans	X58303	348	203	58%
Streptococcus pneumoniae	X58312	349	236	68%
Streptococcus pyogenes	X59029	333	119	36%
Streptomyces acidiscabies	D63865	446	234	52%
Streptomyces albidoflavus	Z76676	426	200	47%
Streptomyces albus	X53163	341	160	47%
Streptomyces ambofaciens	M27245	446	204	46%
Streptomyces bikiniensis	X79851	445	196	44%
Streptomyces bluensis	X79324	445	206	46%
Streptomyces bottropensis	D63868	446	177	40%
Streptomyces brasiliensis	X53162	335	167	50%
Streptomyces caelestis	X80824	445	212	48%
Streptomyces coelicolor	Y00411	441	199	45%
Streptomyces diastaticus	X53161	302	163	54%
Streptomyces diastatochromogenes	D63867	446	212	48%
Streptomyces espinosus	X80826	444	188	42%
Streptomyces eurythermus	D63870	445	211	47%
Streptomyces felleus	Z76681	426	200	47%
Streptomyces galbus	X79325	442	194	44%
Streptomyces glaucescens	X79322	444	287	65%
Streptomyces gougerotii	Z76687	427	184	43%
Streptomyces griseus	X61478	443	173	39%
Streptomyces hygroscopicus	X79853	444	198	45%
Streptomyces intermedius	Z76686	425	208	49%
Streptomyces lavendulae	X53173	318	195	61%
Streptomyces limosus	Z76679	426	200	47%
Streptomyces lincolnensis	X79854	445	183	41%
Streptomyces macrosporus	Z68099	431	198	46%
Streptomyces mashuensis	X79323	441	290	66%
Streptomyces megasporus	Z68100	436	251	58%
Streptomyces neyagawaensis	D63869	446	190	43%
Streptomyces nodosus	AF114033	445	195	44%
Streptomyces odorifer	Z76682	427	200	47%
Streptomyces ornatus	X79326	441	140	32%

Streptomyces pseudogriseolus	X80827	444	287	65%
Streptomyces purpureus	X53170	304	126	41%
Streptomyces rimosus	X62884	438	194	44%
Streptomyces rutgersensis	Z76688	427	184	43%
Streptomyces sampsonii	D63871	446	292	65%
Streptomyces scabiei	D63862	448	180	40%
Streptomyces setonii	D63872	446	200	45%
Streptomyces sp.	D63866	445	214	48%
Streptomyces subrutilus	X80825	444	197	44%
Streptomyces tendae	D63873	445	221	50%
Streptomyces thermodiastaticus	Z68101	428	184	43%
Streptomyces thermolineatus	Z68097	432	200	46%
Streptomyces thermoviolaceus	Z68096	432	198	46%
Streptomyces thermovulgaris	Z68098	436	204	47%
Synechococcus sp. PCC 6301	X03538	429	217	51%
Synechocystis sp. PCC 6803	D64000	432	150	35%
Thermomicrobium roseum	M34115	442	207	47%
Thermotoga maritima	M21774	461	286	62%
Thermus aquaticus	L09663	428	253	59%
Thermus thermophilus	X07998	445	266	60%
Treponema pallidum (rRNA A)	AE001204	455	235	52%
Tropheryma whipplei	X99636	444	249	56%
Ureaplasma urealyticum	AE002112	433	260	60%
Vibrio cholerae	X76337	444	251	57%
Vibrio parahaemolyticus	X56580	417	171	41%
Xanthomonas albilineans	X95918	434	172	40%
Xanthomonas campestris	**NPA**	423	156	37%
Xylella fastidiosa	AE003861	448	170	38%
Yersinia pestis	L37604	423	233	55%
Yersinia pseudotuberculosis	Z21939	434	231	53%

Eukaryotic Chloroplast	**Accession**	**Comp BP**	**PC**	**Acc**
Apodanthes sp	**NPA**	416	198	48%
Astasia longa	X14386	428	193	45%
Babesia bovis	U06105	398	200	50%
Chlamydomonas humicola	AF374186	368	245	67%
Chlamydomonas reinhardtii	J01395	418	231	55%
Chlorella vulgaris	AB001684	431	304	71%
Corethron criophilum	**NPA**	412	189	46%
Cryptomonas sp.	X56805	421	157	37%
Cyanidium caldarium	X52985	421	221	52%
Cyanophora paradoxa	X81840	428	236	55%
Cynomorium coccineum	U67743	401	107	27%
Cytinus ruber	U47845	413	189	46%
Emiliania huxleyi	X82156	413	215	52%
Euglena gracilis	X12890	425	81	19%
Glaucocystis nostochinearum	X82496	427	234	55%

Gloeochaete wittrockiana	X82495	428	208	49%
Heterosigma akashiwo	M82860	428	264	62%
Hydnora africana	U67745	401	232	58%
Marchantia polymorpha	X04465	410	241	59%
Mitrastema yamamotoi	U67742	393	108	27%
Nicotiana tabacum	V00165	414	164	40%
Palmaria palmata	Z18289	411	179	44%
Pilostyles thurberi	U67741	378	154	41%
Plasmodium falciparum (plastid-like)	X57167	409	109	27%
Plasmodium vivax	AF040974	260	99	38%
Polytoma obtusum	AF374187	368	149	40%
Polytoma oviforme	AF374188	370	236	64%
Polytoma uvella	AF374189	437	140	32%
Pylaiella littoralis	X14873	416	126	30%
Ricinus communis	L37580	407	173	43%
Skeletonema pseudocostatum	X82155	413	168	41%
Toxoplasma gondii	U87145	420	182	43%
Zea mays	Z00028	423	173	41%

Eukaryotic Mitochondrion	**Accession**	**Comp BP**	**PC**	**Acc**
Acanthamoeba castellanii	U03732	405	131	32%
Afrixalus fornasini	**NPA**	209	50	24%
Albinaria caerulea	X83390	178	61	34%
Alligator mississippiensis	L28074	208	66	32%
Amblysomus hottentotus	M95108	229	62	27%
Anas platyrhynchos	L16770	233	60	26%
Anopheles gambiae	L20934	198	47	24%
Anopheles quadrimaculatus	L04272	198	53	27%
Antilocapra americana	M55540	246	103	42%
Apis mellifera	L06178	184	34	18%
Artemia franciscana	X69067	169	9	5%
Ascaris suum	X54253	158	21	13%
Aspergillus nidulans	J01393	387	116	30%
Asterina pectinifera	D16387	201	76	38%
Balaenoptera musculus	X72204	232	113	49%
Bos taurus	J01394	229	102	45%
Bufo boreas boreas	**NPA**	217	101	47%
Bufo peltocephalus	**NPA**	214	104	49%
Caenorhabditis elegans	X54252	166	40	24%
Cafeteria roenbergensis	AF193903	417	126	30%
Ceratophrys sp.	**NPA**	203	82	40%
Chlamydomonas eugametos	AF008237	340	166	49%
Chlamydomonas reinhardtii	X54860	272	93	34%
Chondrus crispus	Z30950	370	100	27%
Chorthippus parallelus ESC	X95574	189	45	24%
Chorthippus parallelus NOR	X95575	189	45	24%
Chrysodidymus synuroideus	**NPA**	427	151	35%

Chrysodidymus synuroideus mg	AF222718	443	157	35%
Coscoroba coscoroba	S76216	236	88	37%
Coturnix coturnix	X57245	218	70	32%
Crithidia fasciculata	X02548	55	8	15%
Crossostoma lacustre	M91245	236	100	42%
Cygnus melancoryphus	S76217	239	97	41%
Cyprinus carpio	X61010	228	125	55%
Damaliscus pygargus	M86499	228	46	20%
Daphnia pulex	Z15015	183	34	19%
Dictyostelium discoideum	D16466	353	126	36%
Didelphis virginiana	Z29573	231	82	35%
Drosophila teissieri	X54011	198	40	20%
Drosophila virilis	X05914	222	27	12%
Drosophila yakuba	X03240	198	40	20%
Eleutherodactylus coqui	NPA	184	39	21%
Equus caballus	X79547	229	103	45%
Farfantepenaeus notialis	X84357	209	51	24%
Felis catus	U20753	227	71	31%
Gallus gallus	X52392	223	57	26%
Glycine max	M16859	414	83	20%
Harpactes ardens	U94810	214	59	28%
Harpochytrium sp. JEL94	AY182005	322	119	37%
Herpetomonas megaseliae	U01006	34	3	9%
Homo sapiens	J01415	236	98	42%
Katharina tunicata	U09810	178	9	5%
Latimeria chalumnae	Z21921	212	62	29%
Leishmania tarentolae	M10126	55	12	22%
Locusta migratoria	X80245	178	10	6%
Loxodonta africana	U60182	228	59	26%
Lumbricus terrestris	U24570	172	32	19%
Lutreolina crassicaudata	U33494	238	49	21%
Macropus giganteus	X86941	228	81	36%
Magicicada tredecim	**NPA**	170	17	10%
Marchantia polymorpha	M68292	407	153	38%
Metridium senile	S75445	280	107	38%
Monosiga brevicollis	AF538053	377	185	49%
Mus musculus	J01420	225	89	40%
Muscardinus avellanarius	X84384	230	93	40%
Mytilus edulis	M83756	197	43	22%
Nephroselmis olivacea	AF110138	423	181	43%
Neurospora crassa	L33367	330	71	22%
Ochromonas danica	AF287134	440	126	29%
Oenothera berteriana	X61277	412	76	18%
Oncorhynchus mykiss	L29771	226	67	30%
Ornithorhynchus anatinus	U33498	222	67	30%
Pan troglodytes	D38113	230	57	25%
Paracentrotus lividus	J04815	202	69	34%
Paramecium tetraurelia	K01751	351	55	16%

Pedinomonas minor	AF116775	326	88	27%
Penicillium chrysogenum	Z23072	345	116	34%
Petromyzon marinus	U11880	204	47	23%
Phalanger orientalis	U33496	226	48	21%
Phascogale tapoatafa	U33497	242	70	29%
Phoca vitulina	X63726	225	70	31%
Physarum polycephalum	X75592	398	69	17%
Phytophthora infestans	U17009	429	204	48%
Pichia canadensis	D49702	318	107	34%
Podospora anserina	X14734	347	75	22%
Porphyra purpurea	AF114794	377	126	33%
Protopterus dolloi	L42813	223	72	32%
Prototheca wickerhamii	X15435	441	222	50%
Puma concolor	U33495	225	91	40%
Pylaiella littoralis	X14874	364	87	24%
Rana catesbeiana	X12841	227	82	36%
Rattus norvegicus	J01438	222	81	36%
Reclinomonas americana	AF007261	468	280	60%
Rhizopus stolonifer	**NPA**	384	222	58%
Rhodomonas salina	AF288090	407	107	26%
Saccharomyces cerevisiae	V00704	373	116	31%
Salmo salar	U12143	226	74	33%
Sceloporus undulatus	L28075	215	92	43%
Schizosaccharomyces pombe	X15738	382	152	40%
Scyliorhinus canicula	Y16067	225	74	33%
Secale cereale	Z14059	414	71	17%
Sphenodon punctatus	L28076	208	47	23%
Spizellomyces punctatus	AF404303	324	110	34%
Stenella coeruleoalba	X78169	229	96	42%
Suillus sinuspaulianus	L47584	515	147	29%
Tetrahymena pyriformis	M12714	372	97	26%
Trachemys scripta	L28077	207	49	24%
Triticum aestivum	Z14078	414	71	17%
Trypanosoma brucei	X02547	59	3	5%
Williopsis saturnus var. mrakii	X71392	324	125	39%
Xenopus laevis	M27605	229	93	41%
Zea mays	X00794	426	128	30%

Eukaryotic Nuclear	**Accession**	**Comp BP**	**PC**	**Acc**
Acanthamoeba castellanii	U07413	490	145	30%
Agmasoma penaei	**NPA**	280	121	43%
Ahnfeltia plicata	Z14139	444	190	43%
Alexandrium fundyense	U09048	393	83	21%
Amblyospora sp.	U68474	376	95	25%
Ameson michaelis	L15741	328	50	15%
Androctonus australis	X77908	423	140	33%
Antonospora scoticae	AF024655	364	73	20%

Artemia salina	X01723	423	128	30%
Audouinella dasyae	L26181	448	203	45%
Audouinella hermannii	AF026040	448	205	46%
Aulacoseira ambigua	X85404	441	97	22%
Babesia bigemina	X59604	397	75	19%
Bacillidium sp.	AF104087	354	111	31%
Balamuthia mandrillaris	AF019071	442	149	34%
Balbiania investiens	AF132294	427	168	39%
Bangia sp. (Northwest Territories/NWT)	AF043355	448	134	30%
Bangiopsis subsimplex	AF168627	442	164	37%
Batrachospermum gelatinosum	AF026045	444	152	34%
Batrachospermum macrosporum	AF026048	442	142	32%
Bonnemaisonia hamifera	L26182	442	170	38%
Bostrychia moritziana	AF203893	447	173	39%
Candida albicans	M60302	450	190	42%
Ceramium rubrum	L26183	447	167	37%
Chlorella luteoviridis	X73998	435	150	34%
Chondrus crispus	Z14140	448	202	45%
Compsopogon coeruleus	AF087124	430	116	27%
Corallina officinalis	L26184	455	165	36%
Crossodonthina koreana	Z36893	421	139	33%
Cryptocercus punctulatus	**NPA**	438	135	31%
Cryptococcus neoformans var. neoformans	L05428	448	216	48%
Culicosporella lunata	AF027683	377	146	39%
Cyanophora paradoxa	X68483	446	174	39%
Cymatosira belgica	X85387	441	132	30%
Cyrtohymena citrina	AF164135	442	154	35%
Dixonielloa grisea	L26187	441	151	34%
Drosophila melanogaster	M21017	458	164	36%
Echinococcus granulosus	U27015	417	97	23%
Edhazardia aedis	AF027684	413	130	31%
Encephalitozoon cuniculi	X98467	355	73	21%
Encephalitozoon hellem	AF118143	356	72	20%
Encephalitozoon sp.	L16867	349	72	21%
Endoreticulatus schubergi	L39109	346	91	26%
Engelmanniella mobilis	AF164134	442	185	42%
Enterocytozoonidae gen. sp.	AF201911	333	150	45%
Erythrotrichia carnea	L26189	456	170	37%
Euglypha rotunda	X77692	441	200	45%
Euplotes aediculatus	M14590	433	106	24%
Flabelliforma montana	AJ252962	270	65	24%
Fragaria x ananassa	X15590	444	142	32%
Gastrostyla steinei	AF164133	441	154	35%
Gelidium vagum	L26190	445	169	38%
Genicularia spirotaenia	**NPA**	440	163	37%
Giardia ardeae	Z17210	352	115	33%
Giardia intestinalis	X52949	355	83	23%
Giardia muris	X65063	344	99	29%

Glaucocystis nostochinearum	X70803	445	150	34%
Gloeochaete wittrockiana	X81901	444	155	35%
Glomus intraradices	X58725	394	168	43%
Glugea atherinae	U15987	345	145	42%
Glugea stephani	AF056015	288	130	45%
Gracilariopsis sp.	M33639	448	207	46%
Halymenia plana	U33133	447	152	34%
Hexamita sp.	Z17224	358	105	29%
Hildenbrandia rubra	L19345	449	157	35%
Homo sapiens	K03432	466	142	30%
Ichthyosporidium sp.	L39110	335	98	29%
Janacekia debaisieuxi	AJ252950	360	119	33%
Lecanora dispersa	**NPA**	350	81	23%
Lilioceris lilii	**NPA**	446	200	45%
Loma acerinae	AJ252951	351	117	33%
Mastigamoeba balamuthi	L23799	490	244	50%
Microgemma sp.	AJ252952	343	132	38%
Microsporidium 57864	U90885	355	94	26%
Mus musculus	X00686	448	154	34%
Mytilus edulis	L24489	433	180	42%
Naegleria gruberi	**NPA**	390	60	15%
Nemalion helminthoides	L26196	447	206	46%
Nemalionopsis shawii	AF506272	439	150	34%
Nosema algerae	AF069063	394	114	29%
Nosema apis	U97150	344	98	28%
Nosema necatrix	U11051	322	77	24%
Okanagana utahensis	U06478	457	180	39%
Onychodromus quadricornutus	X53485	440	120	27%
Ophiopholis aculeata	L28056	390	112	29%
Oxytricha granulifera	AF164122	439	141	32%
Oxytricha longa	AF164125	441	182	41%
Palmaria palmata	Z14142	447	203	45%
Paraurostyla weissei	AF164127	441	176	40%
Paruroleptus lepisma	AF164132	440	148	34%
Paulinella chromatophora	X81811	443	171	39%
Placopecten magellanicus	X53899	428	126	29%
Plasmodium falciparum (A gene)	M19172	503	233	46%
Plasmodium falciparum (S gene)	M19173	492	150	30%
Plasmodium vivax (A gene)	U07367	488	217	44%
Plasmodium vivax (O gene)	U93095	505	193	38%
Plasmodium vivax (S gene)	U07368	509	174	34%
Pleistophora hippoglossoideos	AJ252953	361	96	27%
Pleistophora sp.	U10342	329	109	33%
Pleurotricha lanceolota	AF164128	440	128	29%
Plocamiocolax pulvinata	U09618	453	151	33%
Polydispyrenia simulii	AJ252960	333	93	28%
Porphyra miniata	AF175540	444	102	23%
Porphyridium aerugineum	L27635	442	159	36%

Reticulitermes flavipes	**NPA**	447	194	43%
Rhodella maculata	U21217	436	150	34%
Rhodochaete parvula	AF139462	436	153	35%
Rhodogorgon carriebowensis	AF006089	467	173	37%
Rhodymenia leptophylla	U09621	449	176	39%
Saccharomyces cerevisiae	U53879	451	183	41%
Spraguea lophii	AF033197	345	101	29%
Staurastrum sp. M752	X74752	435	160	37%
Strongylocentrotus purpuratus	L28055	390	127	33%
Stylonychia lemnae	AF164124	442	148	33%
Stylonychia mytilus	AF164123	441	156	35%
Thalassiosira eccentrica	X85396	439	170	39%
Thelohania solenopsae	AF031538	384	81	21%
Thelohania sp.	AF031537	256	67	26%
Thorea hispida	AF506273	437	164	38%
Thorea violacea	AF026042	498	182	37%
Thorea violacea	AF506274	437	191	44%
Toxoplasma gondii (P)	X75453	412	154	37%
Trachipleistophora hominis	AJ002605	355	119	34%
Tritrichomonas foetus	M81842	397	156	39%
Uroleptus gallina	AF164130	441	184	42%
Uroleptus pisces	AF164131	441	177	40%
Urostyla grandis	AF164129	437	175	40%
Vairimorpha necatrix	Y00266	338	84	25%
Vairimorpha sp. Argentina	AF031539	303	78	26%
Vavraia culicis	AJ252961	354	149	42%
Visvesvaria acridophagus	AF024658	399	156	39%
Weiseria palustris	AF132544	369	115	31%
Xenopus laevis	X04025	429	135	31%

23S Ribosomal RNA (23S rRNA)	**Accession**	**Comp BP**	**PC**	**Acc**
Archaea				
Archaeoglobus fulgidus	M64487	817	602	74%
Desulfurococcus mobilis	X05480	874	518	59%
Haloarcula marismortui	X13738	816	459	56%
Haloarcula marismortui rrnA	AF034619	816	408	50%
Haloarcula marismortui rrnB	AF034620	818	452	55%
Halobacterium salinarum	X03407	794	483	61%
Halococcus morrhuae	X05481	798	458	57%
Haloferax volcanii	**NPA**	778	460	59%
Methanococcus jannaschii	U67517	879	574	65%
Methanococcus vannielii	X02729	823	511	62%
Methanospirillum hungatei	M81323	785	314	40%
Methanothermobacter thermautotrophicus	X05482	842	338	40%
Sulfolobus acidocaldarius	M67495	861	505	59%
Thermococcus celer	M67497	860	573	67%

Thermofilum pendens	X14835	867	519	60%
Thermoplasma acidophilum	M32298	798	488	61%
Thermoproteus tenax	**NPA**	867	499	58%

Bacteria	**Accession**	**Comp BP**	**PC**	**Acc**
Acinetobacter calcoaceticus	X87280	820	391	48%
Aeromonas hydrophila	X87281	818	480	59%
Aquifex aeolicus	AE000709	851	482	57%
Bacillus anthracis	X64645	808	388	48%
Bacillus sp.	X60981	803	410	51%
Bacillus subtilis	K00637	835	474	57%
Bartonella bacilliformis	L39095	802	409	51%
Bordetella bronchiseptica	X70371	804	369	46%
Bordetella pertussis	X68323	797	267	34%
Borrelia burgdorferi	M88330	842	367	44%
Bradyrhizobium japonicum	Z35330	787	464	59%
Bradyrhizobium japonicum	X71840	792	470	59%
Burkholderia cepacia	X16368	817	436	53%
Burkholderia mallei	Y17183	821	438	53%
Burkholderia pseudomallei	Y17184	821	433	53%
Campylobacter coli	U09611	837	448	54%
Campylobacter jejuni	Z29326	820	398	49%
Chlamydia suis	U68420	793	291	37%
Chlamydia trachomatis	U68443	819	354	43%
Chlamydophila pneumoniae	U76711	791	320	40%
Chlamydophila psittaci	U68447	831	305	37%
Chlorobium limicola	M62805	722	434	60%
Citrobacter freundii	U77928	821	312	38%
Clostridium botulinum A	X65602	808	431	53%
Clostridium botulinum B	M94178	826	453	55%
Clostridium botulinum B	M94259	823	403	49%
Clostridium botulinum E	M94261	819	452	55%
Coxiella burnetii	X79704	773	307	40%
Deinococcus radiodurans	AE002087	809	335	41%
Deinococcus radiodurans	AE001886	810	384	47%
Enterococcus faecalis	X79341	824	432	52%
Erysipelothrix rhusiopathiae	AB019250	817	406	50%
Escherichia coli	J01695	830	458	55%
Flexibacter flexilis	M62806	741	243	33%
Frankia sp.	M55343	875	406	46%
Geobacillus stearothermophilus	K02663	817	519	64%
Haemophilus influenzae (operons A-F)	U32742	820	360	44%
Helicobacter pylori	U27270	785	337	43%
Heliobacterium chlorum	**NPA**	819	472	58%
Klebsiella pneumoniae	X87284	828	423	51%
Lactobacillus delbrueckii	X68426	823	431	52%
Lactococcus lactis	X68434	828	358	43%

Leptospira interrogans	X14249	831	360	43%
Listeria monocytogenes	X64533	827	388	47%
Listeria monocytogenes	X68420	838	505	60%
Micrococcus luteus	X06484	854	415	49%
Mycobacterium leprae	X56657	862	478	55%
Mycobacterium tuberculosis	Z73902	871	375	43%
Mycoplasma genitalium	U39694	832	329	40%
Mycoplasma pneumoniae	X68422	829	413	50%
Myroides odoratus	M62807	701	214	31%
Neisseria gonorrhoeae	X67293	819	340	42%
Neisseria meningitidis	X67300	820	366	45%
Parachlamydia acanthamoebae	Y07555	802	324	40%
Pirellula marina	X07408	790	449	57%
Plesiomonas shigelloides	X65487	762	282	37%
Pseudomonas aeruginosa	Y00432	824	399	48%
Rhodobacter capsulatus	X06485	786	370	47%
Rhodobacter sphaeroides	X53853	788	447	57%
Rhodococcus erythropolis	AF001265	868	428	49%
Rhodopseudomonas palustris	X71839	776	392	51%
Rickettsia prowazekii	AJ235270	788	412	52%
Rickettsia rickettsii	U11022	742	488	66%
Ruminobacter amylophilus	X06765	805	323	40%
Simkania negevensis	U68460	796	344	43%
Staphylococcus aureus	X68425	836	445	53%
Staphylococcus carnosus	X68419	837	476	57%
Streptomyces ambofaciens	M27245	867	516	60%
Streptomyces griseus	X61478	867	354	41%
Synechococcus sp. PCC 6301	X00512	789	448	57%
Thermotoga maritima	M67498	874	541	62%
Thermus aquaticus	**NPA**	820	463	56%
Thermus thermophilus	X12612	805	437	54%
Treponema pallidum (rRNA A)	AE001204	840	281	33%
Tropheryma whipplei	AF190687	828	397	48%

Eukaryotic Chloroplast	**Accession**	**Comp BP**	**PC**	**Acc**
Alnus incana	M75719	772	339	44%
Astasia longa	X14386	826	190	23%
Chlamydomonas eugametos	Z17234	821	282	34%
Chlamydomonas frankii	L43352	770	286	37%
Chlamydomonas geitleri	L43353	785	303	39%
Chlamydomonas gelatinosa	Z15151	775	312	40%
Chlamydomonas humicola	L42989	773	285	37%
Chlamydomonas indica	X68893	818	287	35%
Chlamydomonas iyengarii	L43354	775	360	46%
Chlamydomonas komma	L43502	775	378	49%
Chlamydomonas mexicana	L49148	782	330	42%
Chlamydomonas pallidostigmatica	L43503	816	326	40%

Chlamydomonas peterfii	L43538	779	351	45%
Chlamydomonas pitschmannii	Z15152	782	341	44%
Chlamydomonas reinhardtii	X15727	805	379	47%
Chlamydomonas sp. SAG 66.72	L43539	809	362	45%
Chlamydomonas starrii	L43504	777	323	42%
Chlamydomonas zebra	L43356	774	336	43%
Chlorella ellipsoidea	M36158	772	315	41%
Conopholis americana	X59768	746	179	24%
Epifagus virginiana	X62099	757	257	34%
Euglena gracilis	X12890	805	280	35%
Marchantia polymorpha	X04465	774	288	37%
Nicotiana tabacum	Z00044	782	348	45%
Odontella sinensis	Z67753	777	304	39%
Oryza sativa	X15901	794	308	39%
Palmaria palmata	Z18289	779	345	44%
Pisum sativum	X55033	776	229	30%
Plasmodium falciparum (plastid-like)	X61660	743	157	21%
Toxoplasma gondii (plastid-like)	U18086	693	275	40%
Zea mays	Z00028	814	296	36%

Eukaryotic Mitochondrion	**Accession**	**Comp BP**	**PC**	**Acc**
Acanthamoeba castellanii	U03732	721	240	33%
Aedes albopictus	X01078	307	53	17%
Albinaria caerulea	X83390	242	54	22%
Albinaria turrita	X71393	239	36	15%
Allomyces macrogynus	U41288	735	303	41%
Antilocapra americana	M55540	335	114	34%
Apis mellifera	L06178	298	31	10%
Artemia salina	X12965	258	49	19%
Ascaris suum	X54253	194	41	21%
Aspergillus nidulans	J01390	670	309	46%
Balaenoptera musculus	X72204	329	126	38%
Bos taurus	J01394	332	94	28%
Cacozeliana lacertina	AF101007	305	82	27%
Caenorhabditis elegans	X54252	201	63	31%
Cafeteria roenbergensis	AF193903	685	332	48%
Capra hircus	M55541	298	127	43%
Cepaea nemoralis	U23045	214	67	31%
Cervus unicolor	M35875	296	109	37%
Chlamydomonas eugametos	AF008237	458	160	35%
Chlamydomonas reinhardtii	X54860	438	167	38%
Chondrus crispus	Z46224	661	273	41%
Crithidia fasciculata	X02548	144	2	1%
Crithidia oncopelti	X51736	133	4	3%
Crossostoma lacustre	M91245	332	138	42%
Damaliscus pygargus	M86499	336	125	37%
Dictyostelium discoideum	D16466	673	182	27%

Didelphis virginiana	Z29573	337	110	33%
Drosophila melanogaster	X53506	303	64	21%
Drosophila yakuba	X03240	282	60	21%
Equus caballus	X79547	339	127	37%
Euhadra herklotsi	Z71693	224	62	28%
Gallus gallus	X52392	321	81	25%
Homo sapiens	D38112	321	99	31%
Hydropotes inermis	M35876	297	114	38%
Katharina tunicata	U09810	281	49	17%
Leishmania tarentolae	X02354	142	13	9%
Leptomonas sp.	J03814	134	5	4%
Locusta migratoria	X80245	286	86	30%
Loligo bleekeri	AB009838	289	49	17%
Lumbricus terrestris	U24570	254	79	31%
Marchantia polymorpha	M68929	705	296	42%
Meloidogyne javanica	L76261	139	6	4%
Muntiacus reevesi	M35877	369	103	28%
Mus musculus	J01420	335	93	28%
Mytilus edulis	M83756	288	53	18%
Neurospora crassa	X55443	678	278	41%
Ochromonas danica	AF287134	692	292	42%
Odocoileus virginianus	M35874	303	91	30%
Oenothera berteriana	X02559	668	244	37%
Pan troglodytes	D38113	317	97	31%
Paracentrotus lividus	J04815	326	93	29%
Paracrostoma paludiformis	AF101008	296	79	27%
Paramecium primaurelia	K00634	598	142	24%
Paramecium tetraurelia	K01749	636	211	33%
Pecten maximus	X92688	294	67	23%
Penicillium chrysogenum	D13859	670	215	32%
Phoca vitulina	X63726	328	100	30%
Physarum polycephalum	AF080601	646	177	27%
Pichia canadensis	D31785	672	281	42%
Podospora anserina	X14735	684	263	38%
Porphyra purpurea	AF114794	663	223	34%
Prototheca wickerhamii	X68722	716	437	61%
Pylaiella littoralis	Z48620	668	180	27%
Pyura stolonifera	X74513	206	35	17%
Rana catesbeiana	X12841	293	116	40%
Rattus norvegicus	J01438	274	48	18%
Reclinomonas americana	AF007261	742	306	41%
Rhizopus stolonifer	NPA	676	294	43%
Rhodomonas salina	AF288090	697	247	35%
Saccharomyces cerevisiae	J01527	659	131	20%
Sceloporus undulatus	L28075	307	93	30%
Schizosaccharomyces pombe	X06597	680	287	42%
Strongylocentrotus purpuratus	X12631	246	68	28%
Suillus sinuspaulianus	L47585	689	182	26%

Tetrahymena pyriformis	M58010	667	315	47%
Tetrahymena pyriformis	M58011	657	296	45%
Tragulus napu	M55539	303	128	42%
Triticum aestivum	Z11889	668	201	30%
Trypanosoma brucei	X02547	137	8	6%
Xenopus laevis	M10217	341	123	36%
Zea mays	K01868	677	198	29%

Eukaryotic Nuclear	**Accession**	**Comp BP**	**PC**	**Acc**
Aedes albopictus	L22060	924	441	48%
Arabidopsis thaliana	X52320	930	429	46%
Babesia bigemina	NPA	805	240	30%
Brassica napus	D10840	927	376	41%
Caenorhabditis elegans	X03680	972	514	53%
Candida albicans	L28817	957	379	40%
Chlorella ellipsoidea	D17810	945	379	40%
Crithidia fasciculata	Y00055	859	388	45%
Dictyostelium discoideum	X00601	808	453	56%
Didymium iridis	X60210	950	371	39%
Drosophila melanogaster	M21017	959	472	49%
Encephalitozoon cuniculi	AJ005581	609	228	37%
Entamoeba histolytica	X65163	862	421	49%
Euglena gracilis	X53361	960	201	21%
Filobasidiella neoformans var. bacillispora	L14067	972	536	55%
Fragaria x ananassa	X15589	911	297	33%
Giardia ardeae	X58290	750	278	37%
Giardia intestinalis	X52949	774	257	33%
Giardia muris	X65063	739	268	36%
Herdmania momus	X53538	892	254	28%
Hexamita inflata	**NPA**	723	347	48%
Homo sapiens	J01866	954	279	29%
Lycopersicon esculentum	X13557	944	384	41%
Microsporidium 57864	U90885	617	213	35%
Mucor racemosus	M26190	891	391	44%
Mus musculus	J01871	962	419	44%
Naegleria gruberi	**NPA**	891	460	52%
Nosema apis	U76706	611	253	41%
Nosema apis	U97150	615	205	33%
Oryza sativa	M11585	941	449	48%
Physarum polycephalum	V01159	947	259	27%
Phytophthora megasperma	X75631	962	603	63%
Plasmodium falciparum (A gene)	U21939	998	494	49%
Plasmodium falciparum (S gene)	U48228	1070	439	41%
Plasmodium vivax (A gene)	**NPA**	1010	456	45%
Plasmodium vivax (O gene)	**NPA**	918	389	42%
Plasmodium vivax (S gene)	**NPA**	830	311	37%
Pneumocystis carinii	M86760	966	497	51%

Prorocentrum micans	M14649	870	371	43%
Rattus norvegicus	J01881	716	254	35%
Saccharomyces cerevisiae	U53879	1005	486	48%
Schizosaccharomyces japonicus	Z32848	968	417	43%
Schizosaccharomyces pombe	J01359	993	410	41%
Sinapis alba	X15915	940	372	40%
Tetrahymena thermophila	X54512	939	483	51%
Theileria parva	L28036	841	398	47%
Toxoplasma gondii (P)	X75453	905	443	49%
Trepomonas agilis	AF015455	621	248	40%
Trypanosoma brucei	X05682	853	246	29%
Vairimorpha necatrix	**NPA**	635	264	42%
Xenopus laevis	X59734	958	354	37%
Zea mays	**NPA**	685	275	40%

Appendix D

D.1 Prediction accuracy for 191,994 16S rRNA comparative base pairs grouped by RNA Contact Distance

Columns:

RNA Contact Distance: The number of intervening nucleotides between the 5' and 3' halves of a base pair (Section 2.C.5)

Comp BP Count: Total comparative base pairs observed. Only canonical base pairs (G:C, A:U and G:U) are considered

Comp BP PC Count: Total comparative base pairs predicted correctly

Accuracy: Percentage of total comparative base pairs predicted correctly

Web Reference:

http://www.rna.ccbb.utexas.edu/SIM/4C/mfold_Eval/dist_accuracy_efn2/

RNA Contact Distance	Comp BP Count	Comp BP PC Count	Accuracy
3	1	0	0%
4	533	176	33%
5	6506	4380	67%
6	1954	1020	52%
7	7828	4933	63%
8	3127	1637	52%
9	7041	4283	61%
10	3775	1982	53%
11	6430	3676	57%
12	3615	1977	55%
13	4429	2599	59%
14	3350	1916	57%
15	3710	2236	60%
16	3581	1874	52%
17	3403	1962	58%
18	2965	1675	56%
19	3569	1928	54%
20	2792	1471	53%
21	3540	1772	50%
22	2767	1513	55%
23	2673	1143	43%

24	2185	956	44%
25	2071	817	39%
26	2711	1232	45%
27	2128	937	44%
28	2494	1408	56%
29	2005	958	48%
30	2408	1375	57%
31	2340	811	35%
32	1623	988	61%
33	1589	632	40%
34	1640	1001	61%
35	1666	600	36%
36	1504	899	60%
37	1854	680	37%
38	1129	597	53%
39	1700	661	39%
40	1096	620	57%
41	1370	514	38%
42	1613	892	55%
43	1145	506	44%
44	1655	700	42%
45	1162	550	47%
46	1395	505	36%
47	850	469	55%
48	882	321	36%
49	736	452	61%
50	630	329	52%
51	902	565	63%
52	652	348	53%
53	882	550	62%
54	590	331	56%
55	918	569	62%
56	504	285	57%
57	700	411	59%
58	543	286	53%
59	596	362	61%
60	444	223	50%
61	514	294	57%
62	678	318	47%
63	508	290	57%
64	578	253	44%
65	310	136	44%
66	640	272	43%
67	337	130	39%
68	741	324	44%
69	318	105	33%
70	732	329	45%
71	356	126	35%

72	734	293	40%
73	346	131	38%
74	550	220	40%
75	286	112	39%
76	248	94	38%
77	279	113	41%
78	277	104	38%
79	473	161	34%
80	316	131	41%
81	680	286	42%
82	332	143	43%
83	611	307	50%
84	274	124	45%
85	690	326	47%
86	274	115	42%
87	606	256	42%
88	286	125	44%
89	415	182	44%
90	389	170	44%
91	305	117	38%
92	586	233	40%
93	301	100	33%
94	594	219	37%
95	229	84	37%
96	546	202	37%
97	187	74	40%
98	174	78	45%
99	183	87	48%
100	274	108	39%
101	208	92	44%
102	216	83	38%
103	208	78	38%
104	174	61	35%
105	277	118	43%
106	199	78	39%
107	264	132	50%
108	231	71	31%
109	229	126	55%
110	178	62	35%
111	210	112	53%
112	164	59	36%
113	218	113	52%
114	161	55	34%
115	217	104	48%
116	165	49	30%
117	233	98	42%
118	171	56	33%
119	146	48	33%

120	137	37	27%
121	123	43	35%
122	124	38	31%
123	125	39	31%
124	100	30	30%
125	120	41	34%
126	109	25	23%
127	161	48	30%
128	132	26	20%
129	192	47	24%
130	135	25	19%
131	141	25	18%
132	80	19	24%
133	114	28	25%
134	112	15	13%
135	159	21	13%
136	132	29	22%
137	149	26	17%
138	183	34	19%
139	138	18	13%
140	190	28	15%
141	165	26	16%
142	145	23	16%
143	91	14	15%
144	129	23	18%
145	145	26	18%
146	203	28	14%
147	200	35	18%
148	193	25	13%
149	154	25	16%
150	118	9	8%
151	149	14	9%
152	142	3	2%
153	225	17	8%
154	165	6	4%
155	251	17	7%
156	148	6	4%
157	201	15	7%
158	150	5	3%
159	169	11	7%
160	148	6	4%
161	187	22	12%
162	201	12	6%
163	164	18	11%
164	154	12	8%
165	133	18	14%
166	113	14	12%
167	122	12	10%

168	162	24	15%
169	323	86	27%
170	196	42	21%
171	350	89	25%
172	159	33	21%
173	318	81	25%
174	127	23	18%
175	96	6	6%
176	235	12	5%
177	104	5	5%
178	69	9	13%
179	89	5	6%
180	85	16	19%
181	304	84	28%
182	129	31	24%
183	314	87	28%
184	135	30	22%
185	305	85	28%
186	126	29	23%
187	304	85	28%
188	81	21	26%
189	80	3	4%
190	46	3	7%
191	63	1	2%
192	48	3	6%
193	45	0	0%
194	32	1	3%
195	56	0	0%
196	48	4	8%
197	100	7	7%
198	48	3	6%
199	138	11	8%
200	75	4	5%
201	128	12	9%
202	71	3	4%
203	85	7	8%
204	79	3	4%
205	46	8	17%
206	53	2	4%
207	51	7	14%
208	39	1	3%
209	47	7	15%
210	37	2	5%
211	51	8	16%
212	27	2	7%
213	35	4	11%
214	22	3	14%
215	42	4	10%

216	21	2	10%
217	40	6	15%
218	28	1	4%
219	41	5	12%
220	35	2	6%
221	49	6	12%
222	47	5	11%
223	59	7	12%
224	89	15	17%
225	95	7	7%
226	116	21	18%
227	103	9	9%
228	133	25	19%
229	122	14	11%
230	174	33	19%
231	149	16	11%
232	184	34	18%
233	143	16	11%
234	195	34	17%
235	147	16	11%
236	187	30	16%
237	147	17	12%
238	152	23	15%
239	117	15	13%
240	129	18	14%
241	110	14	13%
242	112	15	13%
243	108	12	11%
244	67	7	10%
245	86	8	9%
246	67	6	9%
247	77	9	12%
248	58	5	9%
249	68	12	18%
250	57	6	11%
251	54	10	19%
252	49	5	10%
253	56	9	16%
254	42	3	7%
255	58	10	17%
256	38	4	11%
257	47	7	15%
258	42	2	5%
259	51	6	12%
260	42	4	10%
261	56	2	4%
262	45	5	11%
263	63	4	6%

264	37	5	14%
265	53	5	9%
266	56	3	5%
267	58	8	14%
268	55	4	7%
269	107	10	9%
270	79	3	4%
271	137	10	7%
272	88	9	10%
273	116	9	8%
274	127	12	9%
275	147	17	12%
276	176	23	13%
277	142	20	14%
278	183	29	16%
279	131	10	8%
280	200	33	17%
281	181	15	8%
282	210	36	17%
283	179	13	7%
284	229	37	16%
285	178	16	9%
286	243	35	14%
287	188	18	10%
288	272	39	14%
289	206	21	10%
290	250	30	12%
291	190	19	10%
292	241	24	10%
293	219	18	8%
294	215	17	8%
295	218	20	9%
296	225	6	3%
297	189	17	9%
298	230	5	2%
299	172	15	9%
300	136	7	5%
301	128	11	9%
302	105	9	9%
303	107	7	7%
304	123	11	9%
305	140	10	7%
306	138	10	7%
307	171	10	6%
308	108	10	9%
309	110	11	10%
310	90	14	16%
311	117	13	11%

312	112	30	27%
313	144	16	11%
314	163	31	19%
315	130	14	11%
316	139	26	19%
317	97	13	13%
318	112	8	7%
319	60	8	13%
320	56	2	4%
321	50	5	10%
322	43	3	7%
323	38	4	11%
324	36	3	8%
325	39	2	5%
326	35	3	9%
327	38	3	8%
328	51	2	4%
329	39	3	8%
330	55	4	7%
331	58	4	7%
332	63	3	5%
333	62	9	15%
334	58	2	3%
335	56	10	18%
336	66	2	3%
337	45	9	20%
338	61	5	8%
339	45	10	22%
340	62	6	10%
341	72	12	17%
342	82	7	9%
343	97	10	10%
344	78	5	6%
345	86	10	12%
346	107	6	6%
347	86	12	14%
348	105	10	10%
349	113	14	12%
350	105	9	9%
351	100	11	11%
352	100	9	9%
353	88	12	14%
354	114	11	10%
355	91	11	12%
356	122	12	10%
357	98	8	8%
358	98	14	14%
359	70	7	10%

360	76	11	14%
361	57	5	9%
362	73	11	15%
363	45	5	11%
364	73	9	12%
365	50	5	10%
366	53	7	13%
367	43	5	12%
368	46	7	15%
369	52	5	10%
370	41	4	10%
371	41	3	7%
372	44	2	5%
373	50	4	8%
374	36	1	3%
375	36	1	3%
376	48	2	4%
377	43	2	5%
378	45	2	4%
379	44	2	5%
380	49	5	10%
381	40	2	5%
382	38	3	8%
383	33	2	6%
384	37	2	5%
385	31	2	6%
386	36	1	3%
387	26	1	4%
388	48	1	2%
389	31	1	3%
390	38	2	5%
391	31	2	6%
392	35	2	6%
393	40	1	3%
394	42	4	10%
395	52	2	4%
396	74	8	11%
397	83	4	5%
398	106	9	8%
399	104	3	3%
400	107	9	8%
401	109	3	3%
402	135	10	7%
403	116	4	3%
404	122	8	7%
405	105	3	3%
406	103	6	6%
407	91	1	1%

408	83	5	6%
409	90	1	1%
410	78	4	5%
411	83	1	1%
412	61	2	3%
413	59	0	0%
414	58	2	3%
415	50	0	0%
416	51	2	4%
417	47	0	0%
418	43	1	2%
419	38	0	0%
420	38	1	3%
421	31	0	0%
422	40	0	0%
423	36	1	3%
424	27	1	4%
425	30	1	3%
426	20	1	5%
427	19	1	5%
428	22	1	5%
429	24	0	0%
430	30	2	7%
431	42	2	5%
432	34	3	9%
433	47	2	4%
434	29	1	3%
435	44	2	5%
436	31	1	3%
437	43	3	7%
438	37	2	5%
439	50	3	6%
440	39	2	5%
441	51	3	6%
442	62	2	3%
443	84	1	1%
444	86	1	1%
445	87	2	2%
446	79	1	1%
447	66	3	5%
448	61	1	2%
449	84	2	2%
450	86	2	2%
451	112	1	1%
452	104	3	3%
453	116	2	2%
454	117	6	5%
455	135	7	5%

456	124	9	7%
457	160	10	6%
458	147	7	5%
459	160	9	6%
460	157	8	5%
461	153	9	6%
462	175	9	5%
463	158	9	6%
464	156	9	6%
465	151	9	6%
466	178	10	6%
467	169	7	4%
468	178	10	6%
469	180	4	2%
470	168	10	6%
471	132	5	4%
472	178	10	6%
473	155	5	3%
474	173	7	4%
475	145	6	4%
476	130	5	4%
477	117	7	6%
478	79	5	6%
479	128	7	5%
480	93	4	4%
481	106	7	7%
482	106	6	6%
483	76	7	9%
484	116	8	7%
485	83	8	10%
486	118	9	8%
487	84	8	10%
488	102	11	11%
489	86	8	9%
490	86	9	10%
491	79	8	10%
492	71	8	11%
493	68	4	6%
494	53	3	6%
495	54	5	9%
496	54	3	6%
497	52	6	12%
498	56	3	5%
499	47	5	11%
500	48	2	4%
501	47	6	13%
502	51	3	6%
503	39	4	10%

504	44	5	11%
505	25	2	8%
506	43	4	9%
507	30	3	10%
508	45	6	13%
509	28	1	4%
510	58	6	10%
511	29	1	3%
512	58	6	10%
513	26	0	0%
514	53	3	6%
515	24	3	13%
516	40	3	8%
517	36	6	17%
518	39	3	8%
519	42	6	14%
520	40	2	5%
521	42	6	14%
522	23	1	4%
523	56	8	14%
524	20	1	5%
525	48	5	10%
526	19	0	0%
527	43	3	7%
528	18	3	17%
529	38	2	5%
530	12	3	25%
531	15	1	7%
532	18	3	17%
533	15	1	7%
534	25	3	12%
535	18	1	6%
536	29	2	7%
537	19	1	5%
538	28	2	7%
539	15	2	13%
540	19	2	11%
541	16	2	13%
542	19	2	11%
543	16	2	13%
544	24	4	17%
545	13	2	15%
546	26	4	15%
547	14	2	14%
548	25	4	16%
549	19	2	11%
550	25	4	16%
551	17	2	12%

552	23	4	17%
553	16	2	13%
554	22	5	23%
555	22	1	5%
556	22	3	14%
557	19	1	5%
558	18	3	17%
559	22	0	0%
560	17	2	12%
561	26	0	0%
562	22	2	9%
563	31	0	0%
564	21	1	5%
565	30	0	0%
566	24	1	4%
567	28	0	0%
568	25	0	0%
569	24	1	4%
570	30	0	0%
571	29	1	3%
572	29	0	0%
573	29	1	3%
574	26	0	0%
575	33	0	0%
576	18	0	0%
577	27	0	0%
578	22	0	0%
579	24	0	0%
580	22	0	0%
581	18	0	0%
582	20	0	0%
583	15	0	0%
584	20	0	0%
585	17	0	0%
586	12	0	0%
587	15	0	0%
588	9	0	0%
589	10	0	0%
590	5	0	0%
591	7	0	0%
592	6	0	0%
593	5	0	0%
594	7	0	0%
595	4	0	0%
596	6	0	0%
597	1	0	0%
598	4	0	0%
599	4	0	0%

600	4	0	0%
601	3	0	0%
602	4	0	0%
603	3	0	0%
604	5	0	0%
605	2	0	0%
606	6	0	0%
607	2	0	0%
608	4	0	0%
609	3	0	0%
610	3	0	0%
611	3	0	0%
612	3	0	0%
613	5	0	0%
614	3	0	0%
615	4	0	0%
616	3	0	0%
617	5	0	0%
618	2	0	0%
619	6	0	0%
620	2	0	0%
621	5	0	0%
622	3	0	0%
623	4	0	0%
624	2	0	0%
625	3	0	0%
626	1	0	0%
627	3	0	0%
628	2	0	0%
629	6	0	0%
630	3	0	0%
631	5	0	0%
632	2	0	0%
633	4	0	0%
634	2	0	0%
635	4	0	0%
636	4	0	0%
637	5	1	20%
638	5	0	0%
639	5	1	20%
640	5	0	0%
641	3	1	33%
642	5	0	0%
643	1	1	100%
644	5	0	0%
645	4	1	25%
646	6	0	0%
647	4	1	25%

648	5	0	0%
649	3	0	0%
650	4	0	0%
651	2	0	0%
652	6	0	0%
653	4	0	0%
654	5	0	0%
655	3	0	0%
656	2	0	0%
657	4	0	0%
658	2	0	0%
659	4	0	0%
660	3	0	0%
661	3	0	0%
662	3	1	33%
663	3	0	0%
664	2	1	50%
665	4	0	0%
666	4	1	25%
667	3	0	0%
668	4	1	25%
669	3	0	0%
670	4	1	25%
671	3	0	0%
672	3	1	33%
673	2	0	0%
674	1	0	0%
675	4	0	0%
676	1	0	0%
677	3	0	0%
678	1	0	0%
679	2	0	0%
680	1	0	0%
681	2	0	0%
682	2	0	0%
683	2	0	0%
684	3	0	0%
685	1	0	0%
686	2	0	0%
688	3	0	0%
689	2	0	0%
690	5	0	0%
691	2	0	0%
692	5	0	0%
693	3	0	0%
694	3	0	0%
695	3	0	0%
696	2	0	0%

697	2	0	0%
698	1	0	0%
699	3	0	0%
700	1	0	0%
701	3	0	0%
703	3	0	0%
705	1	0	0%
706	2	0	0%
707	1	0	0%
708	2	0	0%
709	2	0	0%
710	1	0	0%
711	2	0	0%
712	1	0	0%
714	1	0	0%
715	1	0	0%
716	2	0	0%
717	1	0	0%
718	3	0	0%
719	1	0	0%
720	4	0	0%
721	3	0	0%
722	5	0	0%
723	3	0	0%
724	4	0	0%
725	1	0	0%
726	4	0	0%
727	1	0	0%
728	3	0	0%
730	4	0	0%
731	2	0	0%
732	4	0	0%
733	2	0	0%
734	3	0	0%
735	2	0	0%
736	3	0	0%
737	1	0	0%
738	1	0	0%
740	1	0	0%
742	1	0	0%
746	1	0	0%
748	1	0	0%
750	1	0	0%
752	1	0	0%
754	1	0	0%
756	1	0	0%
759	1	0	0%
761	1	0	0%

763	1	0	0%
764	1	0	0%
766	2	0	0%
768	2	0	0%
770	3	0	0%
772	1	0	0%
774	3	0	0%
776	4	0	0%
778	5	0	0%
780	3	0	0%
782	2	0	0%
783	1	0	0%
784	3	0	0%
785	3	0	0%
786	2	0	0%
787	3	0	0%
788	2	0	0%
789	3	0	0%
790	2	0	0%
791	1	0	0%
792	3	0	0%
793	2	0	0%
794	2	0	0%
795	2	0	0%
796	1	0	0%
797	2	0	0%
799	2	0	0%
800	1	0	0%
801	1	0	0%
802	3	0	0%
803	1	0	0%
804	3	0	0%
805	1	0	0%
806	3	0	0%
807	3	0	0%
808	3	0	0%
809	3	0	0%
810	3	0	0%
811	3	0	0%
812	1	0	0%
813	5	0	0%
814	4	0	0%
815	7	0	0%
816	4	0	0%
817	6	0	0%
818	4	0	0%
819	4	0	0%
820	3	0	0%

821	1	0	0%
822	5	0	0%
823	2	0	0%
824	7	0	0%
825	2	0	0%
826	5	0	0%
827	3	0	0%
828	4	0	0%
829	1	0	0%
830	2	0	0%
831	1	0	0%
832	3	0	0%
833	2	0	0%
834	2	0	0%
835	4	0	0%
836	5	0	0%
837	5	0	0%
838	11	0	0%
839	7	0	0%
840	12	0	0%
841	8	0	0%
842	10	0	0%
843	6	0	0%
844	7	0	0%
845	12	0	0%
846	11	0	0%
847	10	0	0%
848	10	0	0%
849	11	0	0%
850	10	0	0%
851	6	0	0%
852	6	0	0%
853	4	0	0%
854	7	0	0%
855	1	0	0%
856	4	0	0%
857	1	0	0%
858	1	0	0%
859	5	0	0%
860	6	0	0%
861	8	0	0%
862	8	0	0%
863	9	0	0%
864	9	0	0%
865	10	0	0%
866	6	0	0%
867	8	0	0%
868	6	0	0%

869	7	0	0%
870	4	0	0%
871	14	0	0%
872	18	0	0%
873	20	0	0%
874	24	0	0%
875	18	0	0%
876	27	0	0%
877	7	0	0%
878	11	0	0%
879	1	0	0%
880	7	0	0%
881	2	0	0%
882	2	0	0%
883	7	0	0%
884	4	0	0%
885	9	0	0%
886	2	0	0%
887	9	0	0%
888	5	0	0%
889	4	0	0%
890	8	0	0%
891	7	0	0%
892	11	0	0%
893	8	0	0%
894	10	0	0%
895	9	0	0%
896	9	0	0%
897	13	0	0%
898	9	0	0%
899	15	0	0%
900	9	0	0%
901	15	0	0%
902	7	0	0%
903	4	0	0%
904	6	0	0%
905	2	0	0%
906	9	0	0%
907	7	0	0%
908	8	0	0%
909	7	0	0%
910	7	0	0%
911	6	0	0%
912	1	0	0%
914	1	0	0%
916	1	0	0%
917	1	0	0%
918	1	0	0%

919	1	0	0%
920	2	0	0%
921	1	0	0%
922	1	0	0%
924	1	0	0%
927	1	0	0%
928	1	0	0%
929	1	0	0%
930	1	0	0%
931	3	0	0%
932	1	0	0%
933	2	0	0%
934	1	0	0%
935	2	0	0%
936	1	0	0%
938	2	0	0%
940	1	0	0%
942	1	0	0%
947	1	0	0%
949	1	0	0%
951	1	0	0%
965	1	0	0%
966	1	0	0%
967	1	0	0%
968	1	0	0%
969	1	0	0%
970	1	0	0%
989	1	0	0%
991	1	0	0%
993	1	0	0%
1025	1	0	0%
1027	1	0	0%
1029	1	0	0%
1043	1	0	0%
1045	1	0	0%
1046	1	0	0%
1048	1	0	0%
1050	1	0	0%
1052	1	0	0%
1054	1	0	0%
1056	1	0	0%
1079	1	0	0%
1081	1	0	0%
1083	1	0	0%
1087	1	0	0%
1089	1	0	0%
1091	1	0	0%
1092	1	0	0%

1093	2	0	0%
1094	3	0	0%
1095	3	0	0%
1096	7	0	0%
1097	4	0	0%
1098	6	0	0%
1099	3	0	0%
1100	5	0	0%
1101	2	0	0%
1102	2	0	0%
1103	2	0	0%
1104	2	0	0%
1105	1	0	0%
1106	2	0	0%
1107	2	0	0%
1108	1	0	0%
1109	6	0	0%
1110	9	0	0%
1111	9	0	0%
1112	10	0	0%
1113	9	0	0%
1114	10	0	0%
1115	4	0	0%
1116	1	0	0%
1117	1	0	0%
1120	1	0	0%
1121	1	0	0%
1122	1	0	0%
1123	3	0	0%
1124	1	0	0%
1125	4	0	0%
1126	1	0	0%
1127	6	0	0%
1128	4	0	0%
1129	4	0	0%
1130	5	0	0%
1131	5	0	0%
1132	4	0	0%
1133	3	0	0%
1134	3	0	0%
1135	2	0	0%
1136	2	0	0%
1137	3	0	0%
1138	2	0	0%
1139	3	0	0%
1140	1	0	0%
1141	3	0	0%
1142	1	0	0%

1143	4	0	0%
1144	1	0	0%
1145	5	0	0%
1147	4	0	0%
1149	2	0	0%
1153	1	0	0%
1155	1	0	0%
1157	1	0	0%
1173	1	0	0%
1177	1	0	0%
1184	2	0	0%
1186	2	0	0%
1188	2	0	0%
1192	1	0	0%
1194	1	0	0%
1196	2	0	0%
1198	1	0	0%
1200	2	0	0%
1202	1	0	0%
1204	1	0	0%
1214	1	0	0%
1216	1	0	0%
1218	1	0	0%
1223	1	0	0%
1225	2	0	0%
1227	2	0	0%
1229	1	0	0%
1234	1	0	0%
1236	1	0	0%
1238	1	0	0%
1242	1	0	0%
1244	1	0	0%
1245	1	0	0%
1246	1	0	0%
1247	1	0	0%
1249	1	0	0%
1278	1	0	0%
1279	1	0	0%
1280	1	0	0%
1282	1	0	0%
1283	1	0	0%
1291	1	0	0%
1293	1	0	0%
1295	2	0	0%
1297	1	0	0%
1299	1	0	0%
1330	1	0	0%
1332	1	0	0%

1334	1	0	0%
1431	1	0	0%
1433	1	0	0%
1435	1	0	0%
1542	1	0	0%
1544	1	0	0%
1546	1	0	0%
1829	1	0	0%
1831	1	0	0%
1833	1	0	0%

Bibliography

1. Doty, P., et al., *Secondary Structure in Ribonucleic Acids.* Proc Natl Acad Sci U S A, 1959. **45**(4): p. 482-99.

2. Doty, P., et al., *Configurational studies of polynucleotides and ribonucleic acid.* Ann N Y Acad Sci, 1959. **81**: p. 693-708.

3. Fresco, J.R., B.M. Alberts, and P. Doty, *Some molecular details of the secondary structure of ribonucleic acid.* Nature, 1960. **188**: p. 98-101.

4. Moazed, D. and H.F. Noller, *Interaction of tRNA with 23S rRNA in the ribosomal A, P, and E sites.* Cell, 1989. **57**(4): p. 585-97.

5. Ogle, J.M., et al., *Recognition of cognate transfer RNA by the 30S ribosomal subunit.* Science, 2001. **292**(5518): p. 897-902.

6. Ramakrishnan, V., *Ribosome structure and the mechanism of translation.* Cell, 2002. **108**(4): p. 557-72.

7. Moazed, D., J.M. Robertson, and H.F. Noller, *Interaction of elongation factors EF-G and EF-Tu with a conserved loop in 23S RNA.* Nature, 1988. **334**(6180): p. 362-4.

8. Ogle, J.M., et al., *Selection of tRNA by the ribosome requires a transition from an open to a closed form.* Cell, 2002. **111**(5): p. 721-32.

9. Noller, H.F., *tRNA-rRNA interactions and peptidyl transferase.* Faseb J, 1993. **7**(1): p. 87-9.

10. Hansen, J.L., et al., *The structures of four macrolide antibiotics bound to the large ribosomal subunit.* Mol Cell, 2002. **10**(1): p. 117-28.

11. Cech, T.R., *Conserved sequences and structures of group I introns: building an active site for RNA catalysis--a review.* Gene, 1988. **73**(2): p. 259-71.

12. Cech, T.R., *Self-splicing of group I introns.* Annu Rev Biochem, 1990. **59**: p. 543-68.

13. Cate, J.H., et al., *Crystal structure of a group I ribozyme domain: principles of RNA packing.* Science, 1996. **273**(5282): p. 1678-85.

14. Mandal, M., et al., *Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria.* Cell, 2003. **113**(5): p. 577-86.

15. Winkler, W.C., et al., *An mRNA structure that controls gene expression by binding S-adenosylmethionine.* Nat Struct Biol, 2003. **10**(9): p. 701-7.

16. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.

17. Zamore, P.D., et al., *RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.* Cell, 2000. **101**(1): p. 25-33.

18. Fire, A., et al., *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.* Nature, 1998. **391**(6669): p. 806-11.

19. Paddison, P.J., A.A. Caudy, and G.J. Hannon, *Stable suppression of gene expression by RNAi in mammalian cells.* Proc Natl Acad Sci U S A, 2002. **99**(3): p. 1443-8.

20. He, L., et al., *A microRNA polycistron as a potential human oncogene.* Nature, 2005. **435**(7043): p. 828-33.

21. Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature, 2005. **435**(7043): p. 834-8.

22. O'Donnell, K.A., et al., *c-Myc-regulated microRNAs modulate E2F1 expression.* Nature, 2005. **435**(7043): p. 839-43.

23. Hatfield, S.D., et al., *Stem cell division is regulated by the microRNA pathway.* Nature, 2005. **435**(7044): p. 974-8.

24. Zuker, M. and D. Sankoff, *Rna Secondary Structures and Their Prediction.* Bulletin of Mathematical Biology, 1984. **46**(4): p. 591-621.

25. Gutell, R.R., J.C. Lee, and J.J. Cannone, *The accuracy of ribosomal RNA comparative structure models.* Curr Opin Struct Biol, 2002. **12**(3): p. 301-10.

26. Delisi, C. and D.M. Crothers, *Prediction of RNA secondary structure.* Proc Natl Acad Sci U S A, 1971. **68**(11): p. 2682-5.

27. Tinoco, I., Jr., O.C. Uhlenbeck, and M.D. Levine, *Estimation of secondary structure in ribonucleic acids.* Nature, 1971. **230**(5293): p. 362-7.

28. Devoe, H. and I. Tinoco, Jr., *The stability of helical polynucleotides: base contributions.* J Mol Biol, 1962. **4**: p. 500-17.

29. Tinoco, I., Jr., et al., *Improved estimation of secondary structure in ribonucleic acids.* Nat New Biol, 1973. **246**(150): p. 40-1.

30. Uhlenbeck, O.C., et al., *Stability of RNA hairpin loops: A 6 -C m -U 6.* J Mol Biol, 1973. **73**(4): p. 483-96.

31. Borer, P.N., et al., *Stability of ribonucleic acid double-stranded helices.* J Mol Biol, 1974. **86**(4): p. 843-53.

32. Nussinov, R. and A.B. Jacobson, *Fast algorithm for predicting the secondary structure of single-stranded RNA.* Proc Natl Acad Sci U S A, 1980. **77**(11): p. 6309-13.

33. Zuker, M. and P. Stiegler, *Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.* Nucleic Acids Res, 1981. **9**(1): p. 133-48.

34. Jacobson, A.B., et al., *Some simple computational methods to improve the folding of large RNAs.* Nucleic Acids Res, 1984. **12**(1 Pt 1): p. 45-52.

35. Freier, S.M., et al., *Improved free-energy parameters for predictions of RNA duplex stability.* Proc Natl Acad Sci U S A, 1986. **83**(24): p. 9373-7.

36. Jaeger, J.A., D.H. Turner, and M. Zuker, *Improved predictions of secondary structures for RNA.* Proc Natl Acad Sci U S A, 1989. **86**(20): p. 7706-10.

37. Walter, A.E., et al., *Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding.* Proc Natl Acad Sci U S A, 1994. **91**(20): p. 9218-22.

38. Zuker, M., *On finding all suboptimal foldings of an RNA molecule.* Science, 1989. **244**(4900): p. 48-52.

39. Hofacker, I.L., et al., *Fast Folding and Comparison of Rna Secondary Structures.* Monatshefte Fur Chemie, 1994. **125**(2): p. 167-188.

40. Konings, D.A. and R.R. Gutell, *A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs.* Rna, 1995. **1**(6): p. 559-74.

41. Fields, D.S. and R.R. Gutell, *An analysis of large rRNA sequences folded by a thermodynamic method.* Fold Des, 1996. **1**(6): p. 419-30.

42. SantaLucia, J., Jr. and D.H. Turner, *Measuring the thermodynamics of RNA secondary structure formation.* Biopolymers, 1997. **44**(3): p. 309-19.

43. Mathews, D.H., et al., *An updated recursive algorithm for RNA secondary structure prediction with improved thermodynamic parameters.* Molecular Modeling of Nucleic Acids, 1998. **682**: p. 246-257.

44. Mathews, D.H., et al., *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.* J Mol Biol, 1999. **288**(5): p. 911-40.

45. Holley, R.W., et al., *Structure of a Ribonucleic Acid.* Science, 1965. **147**: p. 1462-5.

46. Levitt, M., *Detailed molecular model for transfer ribonucleic acid.* Nature, 1969. **224**(221): p. 759-63.

47. Robertus, J.D., et al., *Structure of yeast phenylalanine tRNA at 3 A resolution.* Nature, 1974. **250**(467): p. 546-51.

48. Suddath, F.L., et al., *Three-dimensional structure of yeast phenylalanine transfer RNA at 3.0angstroms resolution.* Nature, 1974. **248**(443): p. 20-4.

49. Fox, G.W. and C.R. Woese, *5S RNA secondary structure.* Nature, 1975. **256**(5517): p. 505-7.

50. Woese, C.R., et al., *Secondary structure model for bacterial 16S ribosomal RNA: phylogenetic, enzymatic and chemical evidence.* Nucleic Acids Res, 1980. **8**(10): p. 2275-93.

51. Noller, H.F., et al., *Secondary structure model for 23S ribosomal RNA.* Nucleic Acids Res, 1981. **9**(22): p. 6167-89.

52. Gutell, R.R., et al., *Comparative anatomy of 16-S-like ribosomal RNA.* Prog Nucleic Acid Res Mol Biol, 1985. **32**: p. 155-216.

53. Gutell, R.R., et al., *Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods.* Nucleic Acids Res, 1992. **20**(21): p. 5785-95.

54. Woese, C.R., et al., *Detailed analysis of the higher-order structure of 16S-like ribosomal ribonucleic acids.* Microbiol Rev, 1983. **47**(4): p. 621-69.

55. Gutell, R.R., N. Larsen, and C.R. Woese, *Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective.* Microbiol Rev, 1994. **58**(1): p. 10-26.

56. Gutell, R.R., M.N. Schnare, and M.W. Gray, *A compilation of large subunit (23S- and 23S-like) ribosomal RNA structures.* Nucleic Acids Res, 1992. **20 Suppl**: p. 2095-109.

57. Gutell, R.R., M.W. Gray, and M.N. Schnare, *A compilation of large subunit (23S and 23S-like) ribosomal RNA structures: 1993.* Nucleic Acids Res, 1993. **21**(13): p. 3055-74.

58. Gutell, R.R., *Collection of small subunit (16S- and 16S-like) ribosomal RNA structures.* Nucleic Acids Res, 1993. **21**(13): p. 3051-4.

59. Gutell, R.R., *Collection of small subunit (16S- and 16S-like) ribosomal RNA structures: 1994.* Nucleic Acids Res, 1994. **22**(17): p. 3502-7.

60. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.

61. Wimberly, B.T., et al., *Structure of the 30S ribosomal subunit.* Nature, 2000. **407**(6802): p. 327-39.

62. Cannone, J.J., et al., *The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs.* BMC Bioinformatics, 2002. **3**(1): p. 2.

63. Michel, F., A. Jacquier, and B. Dujon, *Comparison of fungal mitochondrial introns reveals extensive homologies in RNA secondary structure.* Biochimie, 1982. **64**(10): p. 867-81.

64. Michel, F. and E. Westhof, *Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis.* J Mol Biol, 1990. **216**(3): p. 585-610.

65. Michel, F., K. Umesono, and H. Ozeki, *Comparative and functional anatomy of group II catalytic introns--a review.* Gene, 1989. **82**(1): p. 5-30.

66. Yu, N., *Comparative Sequence Analysis of Group II Introns, tmRNA, and Database.*, in *Institute for Cellular and Molecular Biology*. 2000, The University of Texas at Austin.

67. James, B.D., et al., *The secondary structure of ribonuclease P RNA, the catalytic element of a ribonucleoprotein enzyme.* Cell, 1988. **52**(1): p. 19-26.

68. Harris, J.K., et al., *New insight into RNase P RNA structure from comparative analysis of the archaeal RNA.* Rna, 2001. **7**(2): p. 220-32.

69. Romero, D.P. and E.H. Blackburn, *A conserved secondary structure for telomerase RNA.* Cell, 1991. **67**(2): p. 343-53.

70. Chen, J.L., M.A. Blasco, and C.W. Greider, *Secondary structure of vertebrate telomerase RNA.* Cell, 2000. **100**(5): p. 503-14.

71. Williams, K.P. and D.P. Bartel, *Phylogenetic analysis of tmRNA secondary structure.* Rna, 1996. **2**(12): p. 1306-10.

72. Guthrie, C. and B. Patterson, *Spliceosomal snRNAs.* Annu Rev Genet, 1988. **22**: p. 387-419.

73. Zwieb, C., *Structure and function of signal recognition particle RNA.* Prog Nucleic Acid Res Mol Biol, 1989. **37**: p. 207-34.

74. Diwa, A., et al., *An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression.* Genes Dev, 2000. **14**(10): p. 1249-60.

75. Grundy, F.J., et al., *Sequence requirements for terminators and antiterminators in the T box transcription antitermination system: disparity between conservation and functional requirements.* Nucleic Acids Res, 2002. **30**(7): p. 1646-55.

76. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2002. **30**(1): p. 17-20.

77. Jaeger, L., F. Michel, and E. Westhof, *Involvement of a GNRA tetraloop in long-range RNA tertiary interactions.* J Mol Biol, 1994. **236**(5): p. 1271-6.

78. Groebe, D.R. and O.C. Uhlenbeck, *Characterization of RNA hairpin loop stability.* Nucleic Acids Res, 1988. **16**(24): p. 11725-35.

79. Tuerk, C., et al., *CUUCGG hairpins: extraordinarily stable RNA secondary structures associated with various biochemical processes.* Proc Natl Acad Sci U S A, 1988. **85**(5): p. 1364-8.

80. Woese, C.R., S. Winker, and R.R. Gutell, *Architecture of ribosomal RNA: constraints on the sequence of "tetra-loops".* Proc Natl Acad Sci U S A, 1990. **87**(21): p. 8467-71.

81. Zuker, M. and A.B. Jacobson, *"Well-determined" regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA.* Nucleic Acids Res, 1995. **23**(14): p. 2791-8.

82. Doshi, K.J., et al., *Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction.* BMC Bioinformatics, 2004. **5**(1): p. 105.

83. Plaxco, K.W., K.T. Simons, and D. Baker, *Contact order, transition state placement and the refolding rates of single domain proteins.* J Mol Biol, 1998. **277**(4): p. 985-94.

84. Diamond, J.M., D.H. Turner, and D.H. Mathews, *Thermodynamics of three-way multibranch loops in RNA*. Biochemistry, 2001. **40**(23): p. 6971-81.

85. Mathews, D.H. and D.H. Turner, *Experimentally derived nearest-neighbor parameters for the stability of RNA three- and four-way multibranch loops.* Biochemistry, 2002. **41**(3): p. 869-80.

86. Mathews, D.H., et al., *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure.* Proc Natl Acad Sci U S A, 2004. **101**(19): p. 7287-92.

87. Chen, G., et al., *Factors affecting thermodynamic stabilities of RNA 3 x 3 internal loops*. Biochemistry, 2004. **43**(40): p. 12865-76.

88. Chen, G. and D.H. Turner, *Consecutive GA pairs stabilize medium-size RNA internal loops*. Biochemistry, 2006. **45**(12): p. 4025-43.

89. Kierzek, E., et al., *Nearest neighbor parameters for Watson-Crick complementary heteroduplexes formed between 2'-O-methyl RNA and RNA oligonucleotides*. Nucleic Acids Res, 2006. **34**(13): p. 3609-14.

90. Lu, Z.J., D.H. Turner, and D.H. Mathews, *A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation.* Nucleic Acids Res, 2006. **34**(17): p. 4912-24.

91. Mathews, D.H. and D.H. Turner, *Prediction of RNA secondary structure by free energy minimization.* Curr Opin Struct Biol, 2006. **16**(3): p. 270-8.

92. Brosius, J., et al., *Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli.* Proc Natl Acad Sci U S A, 1978. **75**(10): p. 4801-5.

93. Carbon, P., et al., *The complete nucleotide sequence of the ribosomal 16-S RNA from Excherichia coli. Experimental details and cistron heterogeneities.* Eur J Biochem, 1979. **100**(2): p. 399-410.

94. Brosius, J., T.J. Dull, and H.F. Noller, *Complete nucleotide sequence of a 23S ribosomal RNA gene from Escherichia coli.* Proc Natl Acad Sci U S A, 1980. **77**(1): p. 201-4.

95. Olsen, G.J., *Comparative analysis of nucleotide sequence data*, in *Health Sciences Center*. 1983, University of Colorado.

96. Weiser, B., R. Gutell, and H. Noller, *XRNA: An X windows environment RNA editing/display package.* Unpublished, 1993.

97. Gutell, R.R., H.F. Noller, and C.R. Woese, *Higher order structure in ribosomal RNA*. Embo J, 1986. **5**(5): p. 1111-3.

98. Gautheret, D., D. Konings, and R.R. Gutell, *G.U base pairing motifs in ribosomal RNA*. Rna, 1995. **1**(8): p. 807-14.

99. Gautheret, D., D. Konings, and R.R. Gutell, *A major family of motifs involving G.A mismatches in ribosomal RNA*. J Mol Biol, 1994. **242**(1): p. 1-8.

100. Gutell, R.R. and C.R. Woese, *Higher order structural elements in ribosomal RNAs: pseudo-knots and the use of noncanonical pairs.* Proc Natl Acad Sci U S A, 1990. **87**(2): p. 663-7.

101. Gautheret, D., S.H. Damberger, and R.R. Gutell, *Identification of base-triples in RNA using comparative sequence analysis.* J Mol Biol, 1995. **248**(1): p. 27-43.

102. Woese, C.R. and R.R. Gutell, *Evidence for several higher order structural elements in ribosomal RNA.* Proc Natl Acad Sci U S A, 1989. **86**(9): p. 3119-22.

103. Stiegler, P., et al., *[Secondary and topographic structure of ribosomal RNA 16S of Escherichia coli]*. C R Seances Acad Sci D, 1980. **291**(12): p. 937-40.

104. Glotz, C., et al., *Secondary structure of the large subunit ribosomal RNA from Escherichia coli, Zea mays chloroplast, and human and mouse mitochondrial ribosomes.* Nucleic Acids Res, 1981. **9**(14): p. 3287-306.

105. Branlant, C., et al., *Primary and secondary structures of Escherichia coli MRE 600 23S ribosomal RNA. Comparison with models of secondary structure for maize chloroplast 23S rRNA and for large portions of mouse and human 16S mitochondrial rRNAs.* Nucleic Acids Res, 1981. **9**(17): p. 4303-24.

106. Zwieb, C., C. Glotz, and R. Brimacombe, *Secondary structure comparisons between small subunit ribosomal RNA molecules from six different species.* Nucleic Acids Res, 1981. **9**(15): p. 3621-40.

107. De Rijk, P., et al., *Database on the structure of large ribosomal subunit RNA.* Nucleic Acids Res, 1994. **22**(17): p. 3495-501.

108. Scott, W.G., J.T. Finch, and A. Klug, *The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage.* Cell, 1995. **81**(7): p. 991-1002.

109. Costa, M. and F. Michel, *Frequent use of the same tertiary motif by self-folding RNAs.* Embo J, 1995. **14**(6): p. 1276-85.

110. Costa, M. and F. Michel, *Rules for RNA recognition of GNRA tetraloops deduced by in vitro selection: comparison with in vivo evolution.* Embo J, 1997. **16**(11): p. 3289-302.

111. Gutell, R.R., et al., *A story: unpaired adenosine bases in ribosomal RNAs.* J Mol Biol, 2000. **304**(3): p. 335-54.

112. Traub, W. and J.L. Sussman, *Adenine-guanine base pairing ribosomal RNA.* Nucleic Acids Res, 1982. **10**(8): p. 2701-8.

113. Elgavish, T., et al., *AA.AG@helix.ends: A:A and A:G base-pairs at the ends of 16 S and 23 S rRNA helices.* J Mol Biol, 2001. **310**(4): p. 735-53.

114. Wimberly, B., *A common RNA loop motif as a docking module and its function in the hammerhead ribozyme.* Nat Struct Biol, 1994. **1**(11): p. 820-7.

115. Leontis, N.B. and E. Westhof, *A common motif organizes the structure of multi-helix loops in 16 S and 23 S ribosomal RNAs.* J Mol Biol, 1998. **283**(3): p. 571-83.

116. Cate, J.H., et al., *RNA tertiary structure mediation by adenosine platforms.* Science, 1996. **273**(5282): p. 1696-9.

117. Nissen, P., et al., *RNA tertiary interactions in the large ribosomal subunit: the A-minor motif.* Proc Natl Acad Sci U S A, 2001. **98**(9): p. 4899-903.

118. Doherty, E.A., et al., *A universal mode of helix packing in RNA.* Nat Struct Biol, 2001. **8**(4): p. 339-43.

119. Klein, D.J., et al., *The kink-turn: a new RNA secondary structure motif.* Embo J, 2001. **20**(15): p. 4214-21.

120. Winkler, W.C., et al., *The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs.* Rna, 2001. **7**(8): p. 1165-72.

121. Gutell, R.R., et al., *Predicting U-turns in ribosomal RNA with comparative sequence analysis.* J Mol Biol, 2000. **300**(4): p. 791-803.

122. Nagaswamy, U. and G.E. Fox, *Frequent occurrence of the T-loop RNA folding motif in ribosomal RNAs.* Rna, 2002. **8**(9): p. 1112-9.

123. Lee, J.C., J.J. Cannone, and R.R. Gutell, *The lonepair triloop: a new motif in RNA structure.* J Mol Biol, 2003. **325**(1): p. 65-83.

124. Lee, J.C., R.R. Gutell, and R. Russell, *The UAA/GAN internal loop motif: a new RNA structural element that forms a cross-strand AAA stack and long-range tertiary interactions.* J Mol Biol, 2006. **360**(5): p. 978-88.

125. Cole, J.R., et al., *The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D169-72.

126. DeSantis, T.Z., et al., *Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB.* Appl Environ Microbiol, 2006. **72**(7): p. 5069-72.

127. Wuyts, J., G. Perriere, and Y. Van De Peer, *The European ribosomal RNA database.* Nucleic Acids Res, 2004. **32**(Database issue): p. D101-3.

128. Griffiths-Jones, S., et al., *Rfam: annotating non-coding RNAs in complete genomes.* Nucleic Acids Res, 2005. **33**(Database issue): p. D121-4.

129. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.

130. Gautheret, D., F. Major, and R. Cedergren, *Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA.* Comput Appl Biosci, 1990. **6**(4): p. 325-31.

131. Laferriere, A., D. Gautheret, and R. Cedergren, *An RNA pattern matching program with enhanced performance and portability.* Comput Appl Biosci, 1994. **10**(2): p. 211-2.

132. Macke, T.J., et al., *RNAMotif, an RNA secondary structure definition and search algorithm.* Nucleic Acids Res, 2001. **29**(22): p. 4724-35.

133. Gautheret, D. and A. Lambert, *Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles.* J Mol Biol, 2001. **313**(5): p. 1003-11.

134. Wilbur, W.J. and D.J. Lipman, *Rapid similarity searches of nucleic acid and protein data banks.* Proc Natl Acad Sci U S A, 1983. **80**(3): p. 726-30.

135. Pearson, W.R. and D.J. Lipman, *Improved tools for biological sequence comparison.* Proc Natl Acad Sci U S A, 1988. **85**(8): p. 2444-8.

136. Pearson, W.R., *Flexible sequence similarity searching with the FASTA3 program package.* Methods Mol Biol, 2000. **132**: p. 185-219.

137. Lin, N., *A System for Identifying RNA Structure Features in Sequence Databases*, in *Institute for Cellular and Molecular Biology*. 2000, University of Texas at Austin.

138. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2007. **35**(Database issue): p. D5-12.

139. Needleman, S.B. and C.D. Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins.* J Mol Biol, 1970. **48**(3): p. 443-53.

140. Sankoff, D., *Matching sequences under deletion-insertion constraints.* Proc Natl Acad Sci U S A, 1972. **69**(1): p. 4-6.

141. Smith, T.F. and M.S. Waterman, *Identification of common molecular subsequences.* J Mol Biol, 1981. **147**(1): p. 195-7.

142. Waterman, M.S., *Introduction to Computational Biology: Maps, Sequences and Genomes (Interdisciplinary Statistics)*. 1995: Chapman & Hall/CRC. 448.

143. Dayhoff, M.O., *Computer analysis of protein evolution.* Sci Am, 1969. **221**(1): p. 86-95.

144. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.

145. Lipman, D.J., S.F. Altschul, and J.D. Kececioglu, *A tool for multiple sequence alignment.* Proc Natl Acad Sci U S A, 1989. **86**(12): p. 4412-5.

146. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.

147. Higgins, D.G. and P.M. Sharp, *CLUSTAL: a package for performing multiple sequence alignment on a microcomputer.* Gene, 1988. **73**(1): p. 237-44.

148. Higgins, D.G., J.D. Thompson, and T.J. Gibson, *Using CLUSTAL for multiple sequence alignments.* Methods Enzymol, 1996. **266**: p. 383-402.

149. Thompson, J.D., et al., *The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.* Nucleic Acids Res, 1997. **25**(24): p. 4876-82.

150. Jeanmougin, F., et al., *Multiple sequence alignment with Clustal X.* Trends Biochem Sci, 1998. **23**(10): p. 403-5.

151. Notredame, C., L. Holm, and D.G. Higgins, *COFFEE: an objective function for multiple sequence alignments.* Bioinformatics, 1998. **14**(5): p. 407-22.

152. Notredame, C., D.G. Higgins, and J. Heringa, *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* J Mol Biol, 2000. **302**(1): p. 205-17.

153. Nuin, P.A., Z. Wang, and E.R. Tillier, *The accuracy of several multiple sequence alignment programs for proteins.* BMC Bioinformatics, 2006. **7**: p. 471.

154. Sakakibara, Y., et al., *Stochastic context-free grammars for tRNA modeling.* Nucleic Acids Res, 1994. **22**(23): p. 5112-20.

155. Eddy, S.R. and R. Durbin, *RNA sequence analysis using covariance models.* Nucleic Acids Res, 1994. **22**(11): p. 2079-88.

156. Matsui, H., K. Sato, and Y. Sakakibara, *Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures.* Bioinformatics, 2005. **21**(11): p. 2611-7.

157. Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc Natl Acad Sci U S A, 1998. **95**(25): p. 14863-8.

158. Lambert, A., et al., *The ERPIN server: an interface to profile-based RNA motif identification.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W160-5.

159. Gutell, R.R., M.N. Schnare, and M.W. Gray, *A compilation of large subunit (23S-like) ribosomal RNA sequences presented in a secondary structure format.* Nucleic Acids Res, 1990. **18 Suppl**: p. 2319-30.

160. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.

161. Waugh, A., et al., *RNAML: a standard syntax for exchanging RNA information.* Rna, 2002. **8**(6): p. 707-17.

Vita

Kishore John Doshi was born in Mt Kisco, NY USA on June 13th, 1974 as the son of Jeanne and Kishore Doshi, and graduated from high school in 1992. He received his B.S. in Chemical Engineering (with honors) from The University of Texas at Austin in 1996. He then joined UOP in Des Plaines, IL USA and entered their career development program where he was engaged in the research and development related to the commercialization of novel technology for creating Linear Alpha Olefins using supercritical ethylene. In 1998, Kishore joined the Computer Aided Schedule A (CASA) software development team at UOP as an engineering liaison and was accepted into the Computer Science graduate program at DePaul University. CASA was a novel client/server software tool suite for collaborative engineering. In 1999, Kishore was promoted to Senior Software Engineer and contributed significantly to CASA. In 2001 he received his M.S. in Computer Science (with distinction) and left UOP to join Cap Gemini Ernst and Young in Chicago, IL USA as a Senior Software Consultant. At Cap Gemini Ernst and Young, Kishore mentored junior developers and led several, successful full lifecycle software development projects. He left Cap Gemini Ernst and Young and entered the Molecular Biology graduate program at the University of Texas at Austin in September 2001.

Publications:

Doshi K.J., Cannone J.J., Cobaugh C.W. and Gutell R.R (2004). Evaluation of the suitability of free energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*. **5**:105.

Permanent address: 14928 Bescott Dr, Austin TX 78728

This dissertation was typed by the author.